

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 979 182**

51 Int. Cl.:

G11C 13/00 (2006.01)
G16B 30/00 (2009.01)
G16B 30/20 (2009.01)
G16B 50/00 (2009.01)
G06N 3/12 (2013.01)
G06N 3/123 (2013.01)
H03M 7/40 (2006.01)
C12N 9/22 (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

- 86 Fecha de presentación y número de la solicitud internacional: **16.11.2017 PCT/US2017/062098**
87 Fecha y número de publicación internacional: **24.05.2018 WO18094108**
96 Fecha de presentación y número de la solicitud europea: **16.11.2017 E 17872172 (6)**
97 Fecha y número de publicación de la concesión europea: **15.05.2024 EP 3542294**

54 Título: **Almacenamiento de datos basado en ácidos nucleicos**

30 Prioridad:

16.11.2016 US 201662423058 P
09.02.2017 US 201762457074 P
02.03.2017 US 201762466304 P

45 Fecha de publicación y mención en BOPI de la
traducción de la patente:
24.09.2024

73 Titular/es:

CATALOG TECHNOLOGIES, INC. (100.0%)
529 Main Street, Suite 127
Charlestown, MA 02129, US

72 Inventor/es:

ROQUET, NATHANIEL;
PARK, HYUNJUN y
BHATIA, SWAPNIL, P.

74 Agente/Representante:

FERNÁNDEZ POU, Felipe

ES 2 979 182 T3

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

Almacenamiento de datos basado en ácidos nucleicos

5 Referencia cruzada

Esta solicitud reivindica el beneficio de la Solicitud de Patente Provisional de EE. UU. No. de Serie 62/423,058, presentada el 16 de noviembre de 2016, Solicitud de Patente Provisional de EE. UU. No. de Serie 62/457,074, presentada el 9 de febrero de 2017, y Solicitud de Patente Provisional de EE. UU. No. de Serie 62/466,304, presentada el 2 de marzo de 2017.

Antecedentes

15 El almacenamiento de datos digitales en ácidos nucleicos es un enfoque estable para codificar y almacenar información durante largos períodos de tiempo, con datos almacenados a densidades más altas que los sistemas de almacenamiento en cinta magnética o disco duro. Además, los datos digitales almacenados en moléculas de ácidos nucleicos que se almacenan en condiciones frías y secas pueden recuperarse tan tarde como 60.000 años después o más.

20 Para acceder a datos digitales almacenados en moléculas de ácidos nucleicos, se pueden secuenciar las moléculas de ácidos nucleicos. Como tal, el almacenamiento de datos digitales en ácidos nucleicos puede ser un método ideal para almacenar datos a los que no se accede con frecuencia, pero que pueden tener un gran volumen de información para almacenar o archivar durante largos períodos de tiempo.

25 Los métodos actuales se basan en codificar la información digital (p. ej., código binario) en secuencias de ácidos nucleicos base por base, de modo que la relación base a base en la secuencia se traduzca directamente en la información digital (p. ej., código binario). La secuenciación de datos digitales almacenados en secuencias base por base que pueden leerse en corrientes de bits o bytes de información codificada digitalmente puede ser propensa a errores y costosa de codificar, ya que el costo de la síntesis de novo de ácidos nucleicos base por base puede ser elevado. Las oportunidades para nuevos métodos de realizar el almacenamiento de datos digitales en ácidos nucleicos pueden proporcionar enfoques para codificar y recuperar datos que sean menos costosos y más fáciles de implementar comercialmente.

35 La Publicación de Patente Internacional No. WO2004/009844 se refiere al uso de codones de ácidos nucleicos para transportar información. Bomholt, et al. ("A DNA-Based Archival Storage System") se refiere a un sistema de almacenamiento de archivos basado en ADN que utiliza codificación base por base. Yazdi, et al. ("A Rewritable, Random-Access DNA-Based Storage System") se refiere a una arquitectura de almacenamiento basada en ADN basada en secuencias de ADN dotadas de cadenas de direcciones especializadas. De Silva, et al. ("New Trends of Digital Data Storage in DNA") revisa diversos esquemas de almacenamiento de datos basados en ADN.

Compendio

45 Según un aspecto, se proporciona un método para escribir información en secuencias de ácidos nucleicos según la reivindicación 1.

50 Los aspectos y ventajas adicionales de la presente divulgación resultarán fácilmente evidentes para los expertos en esta técnica a partir de la siguiente descripción detallada, en donde solo se muestran y describen realizaciones ilustrativas de la presente divulgación. Como se comprenderá, la presente divulgación es capaz de otras realizaciones diferentes, y sus diversos detalles son susceptibles de modificaciones en varios aspectos obvios, todo ello sin apartarse de la divulgación. Por consiguiente, los dibujos y descripción deben considerarse como ilustrativos en naturaleza, y no como restrictivos.

Breve descripción de los dibujos

55 En las reivindicaciones adjuntas se establecen con particularidad las características novedosas de la invención. Se obtendrá una mejor comprensión de las características y ventajas de la presente invención por referencia a la siguiente descripción detallada que establece realizaciones ilustrativas, en las que se utilizan los principios de la invención, y los dibujos adjuntos (también "Figura" y "FIG." en la presente memoria), de las que:

60 La FIG. 1 ilustra esquemáticamente una visión general de un proceso para codificar, escribir, acceder, leer y decodificar información digital almacenada en secuencias de ácidos nucleicos;

65 Las FIG. 2A y 2B ilustran esquemáticamente un método de ejemplo de codificación de datos digitales, denominados "datos en la dirección", usando objetos o identificadores (p. ej., moléculas de ácidos nucleicos); la FIG. 2A ilustra la combinación de un objeto de rango (u objeto de dirección) con un objeto de valor de bit (u

objeto de datos) para crear un identificador; la FIG. 2B ilustra una realización del método de datos en dirección en donde los objetos de rango y los objetos de valor de bit son en sí mismos concatenaciones combinatorias de

5 Las FIG. 2A y 2B ilustran esquemáticamente un método de ejemplo de codificación de datos digitales, denominados "datos en la dirección", usando objetos o identificadores (p. ej., moléculas de ácidos nucleicos); la FIG. 2A ilustra la combinación de un objeto de rango (u objeto de dirección) con un objeto de valor de bit (u objeto de datos) para crear un identificador; la FIG. 2B ilustra una realización del método de datos en dirección en donde los objetos de rango y los objetos de valor de bit son en sí mismos concatenaciones combinatorias de otros objetos;

10 Las FIG. 3A y 3B ilustran esquemáticamente un método de ejemplo para codificar información digital usando objetos o identificadores (p. ej., secuencias de ácidos nucleicos); la FIG. 3A ilustra la codificación de información digital usando un objeto de rango como identificador; la FIG. 3B ilustra una realización del método de codificación en donde los objetos de dirección son en sí mismos concatenaciones combinatorias de otros objetos;

15 La FIG. 4 muestra un gráfico de contorno, en espacio log, de una relación entre el espacio combinatorio de posibles identificadores (C, eje x) y el número promedio de identificadores (k, eje y) que pueden construirse para almacenar información de un determinado tamaño (líneas de contorno);

20 La FIG. 5 ilustra esquemáticamente una visión general de un método para escribir información en secuencias de ácidos nucleicos (p. ej., ácido desoxirribonucleico);

25 Las FIG. 6A y 6B ilustran un método de ejemplo, denominado "esquema de producto", para construir identificadores (p. ej., moléculas de ácidos nucleicos) ensamblando combinatoriamente componentes distintos (p. ej., secuencias de ácidos nucleicos); la FIG. 6A ilustra la arquitectura de identificadores construidos usando el esquema de producto; la FIG. 6B ilustra un ejemplo del espacio combinatorio de identificadores que puede construirse usando el esquema de producto;

30 La FIG. 7 ilustra esquemáticamente el uso de la reacción en cadena de la polimerasa de extensión superpuesta para construir identificadores (p. ej., moléculas de ácidos nucleicos) a partir de componentes (p. ej., secuencias de ácidos nucleicos);

35 La FIG. 8 ilustra esquemáticamente el uso de ligación de extremos cohesivos para construir identificadores (p. ej., moléculas de ácidos nucleicos) a partir de componentes (p. ej., secuencias de ácidos nucleicos);

40 La FIG. 9 ilustra esquemáticamente el uso del ensamblaje por recombinasa para construir identificadores (p. ej., moléculas de ácidos nucleicos) a partir de componentes (p. ej., secuencias de ácidos nucleicos);

45 Las FIG. 10A y 10B demuestran la ligación dirigida por molde; la FIG. 10A ilustra esquemáticamente el uso de ligación dirigida por molde para construir identificadores (p. ej., moléculas de ácidos nucleicos) a partir de componentes (p. ej., secuencias de ácidos nucleicos); la FIG. 10B muestra un histograma de los números de copias (abundancias) de 256 secuencias de ácidos nucleicos distintas, cada una de las cuales se ensambló combinatoriamente a partir de seis secuencias de ácidos nucleicos (p. ej., componentes) en una reacción de ligación dirigida por molde combinada;

50 Las FIG. 11A-11G ilustran esquemáticamente un método de ejemplo, denominado "esquema de permutación", para construir identificadores (p. ej., moléculas de ácidos nucleicos) con componentes permutados (p. ej., secuencias de ácidos nucleicos); la FIG. 11A ilustra la arquitectura de identificadores construidos usando el esquema de permutación; la FIG. 11B ilustra un ejemplo del espacio combinatorio de identificadores que se puede construir usando el esquema de permutación; la FIG. 11C muestra una implementación de ejemplo del esquema de permutación con ligación dirigida por molde; la FIG. 11D muestra un ejemplo de cómo la implementación de la FIG. 11C puede modificarse para construir identificadores con componentes permutados y repetidos; la FIG. 11E muestra cómo la implementación de ejemplo de la FIG. 11D puede dar lugar a subproductos no deseados que pueden eliminarse con la selección del tamaño del ácido nucleico; la FIG. 11F muestra otro ejemplo de cómo usar ligación dirigida por molde y selección de tamaño para construir identificadores con componentes permutados y repetidos; la FIG. 11G muestra un ejemplo de cuándo la selección de tamaño puede no lograr aislar un identificador particular de subproductos no deseados;

60 Las FIG. 12A - 12D ilustran esquemáticamente un método de ejemplo, denominado esquema "MchooseK", para construir identificadores (p. ej., moléculas de ácidos nucleicos) con cualquier número, K , de componentes ensamblados (p. ej., secuencias de ácidos nucleicos) a partir de un número mayor, M , de posibles componentes; la FIG. 12A ilustra la arquitectura de identificadores construidos usando el esquema MchooseK; la FIG. 12B ilustra un ejemplo del espacio combinatorio de identificadores que se puede construir usando el esquema MchooseK; la FIG. 12C muestra una implementación de ejemplo del esquema MchooseK usando

ligación dirigida por molde; la FIG. 12D muestra cómo la implementación de ejemplo de la FIG. 12C puede dar lugar a subproductos no deseados que pueden eliminarse con la selección del tamaño del ácido nucleico;

5 Las FIG. 13A y 13B ilustran esquemáticamente un método de ejemplo, denominado "esquema de partición" para construir identificadores con componentes divididos; la FIG. 13A muestra un ejemplo del espacio combinatorio de identificadores que se pueden construir usando el esquema de partición; la FIG. 13B muestra una implementación de ejemplo del esquema de partición usando ligación dirigida por molde;

10 Las FIG. 14A y 14B ilustran esquemáticamente un método de ejemplo, denominado esquema de "cadena sin restricciones" (o USS), para construir identificadores formados por cualquier cadena de componentes a partir de varios componentes posibles; la FIG. 14A muestra un ejemplo del espacio combinatorio de identificadores que se pueden construir usando el esquema USS; la FIG. 14B muestra una implementación de ejemplo del esquema USS usando ligación dirigida por molde;

15 Las FIG. 15A y 15B ilustran esquemáticamente un método de ejemplo, denominado "eliminación de componentes" para construir identificadores eliminando componentes de un identificador parental; la FIG. 15A muestra un ejemplo del espacio combinatorio de identificadores que se puede construir usando el esquema de eliminación de componentes; la FIG. 15B muestra una implementación de ejemplo del esquema de eliminación de componentes usando escisión y reparación dirigidas a doble cadena;

20 La FIG. 16 ilustra esquemáticamente un identificador parental con sitios de reconocimiento de recombinasas donde se pueden construir identificadores adicionales aplicando recombinasas al identificador parental;

25 Las FIG. 17A - 17C ilustran esquemáticamente una visión general de métodos de ejemplo para acceder a porciones de información almacenadas en secuencias de ácidos nucleicos accediendo a un número de identificadores particulares desde un número mayor de identificadores; la FIG. 17A muestra métodos de ejemplo para usar la reacción en cadena de la polimerasa, sondas etiquetadas por afinidad y sondas dirigidas a la degradación para acceder a identificadores que contienen un componente especificado; la FIG. 17B muestra métodos de ejemplo para usar la reacción en cadena de la polimerasa para realizar operaciones 'O' o 'Y' para acceder a identificadores que contienen múltiples componentes especificados; la FIG. 17C muestra métodos de ejemplo para usar etiquetas de afinidad para realizar operaciones 'O' o 'Y' para acceder a identificadores que contienen múltiples componentes especificados;

35 Las FIG. 18A y 18B muestran ejemplos de codificación, escritura y lectura de datos codificados en moléculas de ácidos nucleicos; la FIG. 18A muestra un ejemplo de codificación, escritura y lectura de 5.856 bits de datos; la FIG. 18b muestra un ejemplo de codificación, escritura y lectura de 62.824 bits de datos; y

La FIG. 19 muestra un sistema informático que está programado o configurado de otro modo para implementar los métodos proporcionados en la presente memoria.

40 Descripción detallada

45 Si bien en la presente memoria se han mostrado y descrito diversas realizaciones de la invención, será obvio para los expertos en la técnica que dichas realizaciones se proporcionan únicamente a modo de ejemplo. A los expertos en la técnica se les pueden ocurrir numerosas variaciones, cambios y sustituciones sin alejarse de la invención. Debe entenderse que se pueden emplear diversas alternativas a las realizaciones de la invención descritas en la presente memoria.

50 El término "símbolo", como se usa en la presente memoria, generalmente se refiere a una representación de una unidad de información digital. La información digital puede dividirse o traducirse en una cadena de símbolos. En un ejemplo, un símbolo puede ser un bit y el bit puede tener un valor de '0' o '1'.

55 El término "distintivo", o "único", como se usa en la presente memoria, generalmente se refiere a un objeto que es distinguible de otros objetos en un grupo. Por ejemplo, una secuencia de ácido nucleico distintiva, o única, puede ser una secuencia de ácido nucleico que no tiene la misma secuencia que cualquier otra secuencia de ácido nucleico. Una molécula de ácido nucleico distintiva, o única, puede no tener la misma secuencia que cualquier otra molécula de ácido nucleico. La secuencia o molécula de ácido nucleico distintiva, o única, puede compartir regiones de similitud con otra secuencia o molécula de ácido nucleico.

60 El término "componente", como se usa en la presente memoria, generalmente se refiere a una secuencia de ácido nucleico. Un componente puede ser una secuencia de ácido nucleico distinta. Un componente puede concatenarse o ensamblarse con uno o más de otros componentes para generar otras secuencias o moléculas de ácidos nucleicos.

65 El término "capa", como se usa en la presente memoria, generalmente se refiere a un grupo o combinado de componentes. Cada capa puede comprender un conjunto de componentes distintos de modo que los

componentes de una capa sean diferentes de los componentes de otra capa. Se pueden ensamblar componentes de una o más capas para generar uno o más identificadores.

5 El término "identificador", como se usa en la presente memoria, generalmente se refiere a una molécula de ácido nucleico o una secuencia de ácido nucleico que representa la posición y el valor de una cadena de bits dentro de una cadena de bits más grande. De manera más general, un identificador puede referirse a cualquier objeto que represente o corresponda a un símbolo en una cadena de símbolos. En algunas realizaciones, los identificadores pueden comprender uno o múltiples componentes concatenados.

10 El término "espacio combinatorio", como se usa en la presente memoria, generalmente se refiere al conjunto de todos los identificadores distintos posibles que pueden generarse a partir de un conjunto de objetos de partida, tales como componentes, y un conjunto permisible de reglas sobre cómo modificar esos objetos para formar identificadores. El tamaño de un espacio combinatorio de identificadores creado ensamblando o concatenando componentes puede depender del número de capas de componentes, del número de componentes en cada capa y del método de ensamblaje particular utilizado para generar los identificadores.

15 El término "rango de identificadores", como se usa en la presente memoria, generalmente se refiere a una relación que define el orden de los identificadores en un conjunto.

20 El término "biblioteca de identificadores", como se usa en la presente memoria, generalmente se refiere a una colección de identificadores que corresponden a los símbolos en una cadena de símbolos que representa información digital. En algunas realizaciones, la ausencia de un identificador determinado en la biblioteca de identificadores puede indicar un valor de símbolo en una posición particular. Se pueden combinar una o más bibliotecas de identificadores en un combinado, grupo o conjunto de identificadores. Cada biblioteca de identificadores puede incluir un código de barras único que identifica la biblioteca de identificadores.

30 El término "ácido nucleico", como se usa en la presente memoria, en general se refiere al ácido desoxirribonucleico (ADN), al ácido ribonucleico (ARN), o una variante de los mismos. Un ácido nucleico puede incluir una o más subunidades seleccionadas de adenosina (A), citosina (C), guanina (G), timina (T) y uracilo (U), o variantes de los mismos. Un nucleótido puede incluir A, C, G, T o U, o variantes de los mismos. Un nucleótido puede incluir cualquier subunidad que pueda incorporarse a una cadena de ácido nucleico en crecimiento. Dicha subunidad puede ser A, C, G, T o U, o cualquier otra subunidad que pueda ser específica para una o más complementarias A, C, G, T o U, o complementaria de una purina (es decir, A o G, o variante de la misma) o pirimidina (es decir, C, T o U, o variante de la misma). En algunos ejemplos, un ácido nucleico puede ser monocatenario o bicatenario, en algunos casos, un ácido nucleico es circular.

40 Los términos "molécula de ácido nucleico" o "secuencia de ácido nucleico", como se usan en la presente memoria, generalmente se refieren a una forma polimérica de nucleótidos, o polinucleótido, que puede tener varias longitudes, ya sea desoxirribonucleótidos (ADN) o ribonucleótidos (ARN), o análogos de los mismos. El término "secuencia de ácido nucleico" puede referirse a la representación alfabética de un polinucleótido; alternativamente, el término puede aplicarse al polinucleótido físico mismo. Esta representación alfabética puede introducirse en bases de datos en una computadora que tiene una unidad central de procesamiento y usarse para mapear secuencias de ácido nucleico o moléculas de ácido nucleico en símbolos, o bits, que codifican información digital. Las secuencias de ácidos nucleicos u oligonucleótidos pueden incluir uno o más nucleótidos no estándar, análogos de nucleótidos y/o nucleótidos modificados.

50 Un "oligonucleótido", como se usa en la presente memoria, generalmente se refiere a una secuencia de ácido nucleico monocatenario, y está compuesto típicamente por una secuencia específica de cuatro bases de nucleótidos: adenina (A); citosina (C); guanina (G) y timina (T) o uracilo (U) cuando el polinucleótido es ARN.

Los ejemplos de nucleótidos modificados incluyen, pero no se limitan a, diaminopurina, 5-fluorouracilo, 5-bromouracilo, 5-clorouracilo, 5-yodouracilo, hipoxantina, xantina, 4-acetilcitosina, 5-(carboxihidroximetil)uracilo, 5-carboximetilaminometil-2-tiouridina, 5-carboximetilaminometiluracilo, dihidouracilo, beta-D-galactosilqueosina, inosina, N6-isopenteniladenina, 1-metilguanina, 1-metilinosina, 2,2-dimetilguanina, 2-metiladenina, 2-metilguanina, 3-metilcitosina, 5-metilcitosina, N6-adenina, 7-metilguanina, 5-metilaminometiluracilo, 5-metoxiaminometil-2-tiouracilo, beta-D-manosilqueosina, 5'-metoxycarboximetiluracilo, 5-metoxiuracilo, 2-metil-2-tiouracilo, ácido uracil-5-oxiacético (v), wibutoxosina, pseudouracilo, queosina, 2-tiocitosina, 5-metil-2-tiouracilo, 2-tiouracilo, 4-tiouracilo, 5-metiluracilo, ester metílico del ácido uracil-5-oxiacético, ácido uracil-5-oxiacético (v), 5-metil-2-tiouracilo, 3-(3-amino-3-N-2-carboxipropil)uracilo, (acp3)w, 2,6-diaminopurina y similares. Las moléculas de ácido nucleico también pueden modificarse en el resto de la base (p. ej., en uno o más átomos que típicamente están disponibles para formar un enlace de hidrógeno con un nucleótido complementario y/o en uno o más átomos que típicamente no son capaces de formar un enlace de hidrógeno con un nucleótido complementario), resto de azúcar o cadena principal de fosfato. Las moléculas de ácido nucleico también pueden contener grupos modificados con amina, como aminoalil-dUTP (aa-dUTP) y aminohexilacrilamida-dCTP (aha-dCTP) para permitir la unión covalente de restos reactivos de amina, como los ésteres de N-hidroxisuccinimida (NHS).

El término "cebador", como se usa en la presente memoria, generalmente se refiere a una cadena de ácido nucleico que sirve como punto de partida para la síntesis de ácidos nucleicos, tal como reacción en cadena de la polimerasa (PCR). En un ejemplo, durante la replicación de una muestra de ADN, una enzima que cataliza la replicación inicia la replicación en el extremo 3' de un cebador unido a la muestra de ADN y copia la cadena opuesta.

El término "polimerasa" o "enzima polimerasa", como se usa en la presente memoria, generalmente se refiere a cualquier enzima capaz de catalizar una reacción de polimerasa. Los ejemplos de polimerasas incluyen, sin limitación, una polimerasa de ácido nucleico. La polimerasa puede ser natural o sintetizada. Un ejemplo de polimerasa es una polimerasa $\Phi 29$ o un derivado de la misma. En algunos casos, se usa una transcriptasa o una ligasa (es decir, enzimas que catalizan la formación de un enlace) junto con polimerasas o como alternativa a las polimerasas para construir nuevas secuencias de ácidos nucleicos. Los ejemplos de polimerasas incluyen una ADN polimerasa, una ARN polimerasa, una polimerasa termoestable, una polimerasa de tipo salvaje, una polimerasa modificada, ADN polimerasa I de E. coli, ADN polimerasa T7, ADN polimerasa T4 del bacteriófago $\Phi 29$ ($\phi 29$), ADN polimerasa Taq, polimerasa Tth, polimerasa Tli, polimerasa Pfu polimerasa Pwo, polimerasa VENT, polimerasa DEEPVENT, polimerasa Ex-Taq, polimerasa LA-Taw, polimerasa Sso polimerasa Poc, polimerasa Pab, polimerasa Mth polimerasa ES4, polimerasa Tru, polimerasa Tac, polimerasa Tne, polimerasa Tma, polimerasa Tca, polimerasa Tih, polimerasa Tfi, polimerasa Platinum Taq, polimerasa Tbr, polimerasa Tfl, polimerasa Pfutubo, polimerasa Pyrobest, polimerasa KOD, polimerasa Bst, polimerasa Sac, polimerasa de fragmento Klenow con actividad exonucleasa de 3' a 5', y variantes, productos modificados y derivados de las mismas.

La información digital, tal como datos informáticos, en forma de código binario puede comprender una secuencia o cadena de símbolos. Un código binario puede codificar o representar texto o instrucciones de un procesador de computadora usando, por ejemplo, un sistema numérico binario que tiene dos símbolos binarios, típicamente 0 y 1, denominados bits. La información digital se puede representar en forma de código no binario que puede comprender una secuencia de símbolos no binarios. Cada símbolo codificado puede reasignarse a una cadena de bits única (o "byte"), y la cadena de bits única o byte puede disponerse en cadenas de bytes o corrientes de bytes. Un valor de bit para un bit determinado puede ser uno de dos símbolos (p. ej., 0 o 1). Un bit, que puede comprender una cadena de N bits, puede tener un total de 2^N valores de byte únicos. Por ejemplo, un byte que comprende 8 bits puede producir un total de 2^8 o 256 posibles valores de bytes únicos, y cada uno de los 256 bytes puede corresponder a uno de los 256 posibles símbolos, letras o instrucciones distintas que pueden codificarse con los bytes. Los datos sin procesar (p. ej., archivos de texto e instrucciones de computadora) se pueden representar como cadenas de bytes o corrientes de bytes. Los archivos zip, o archivos de datos comprimidos que comprenden datos sin procesar, también se pueden almacenar en corrientes de bytes, estos archivos se pueden almacenar como corrientes de bytes en forma comprimida y luego descomprimirlos en datos sin procesar antes de ser leídos por la computadora.

Los métodos y sistemas de la presente divulgación se pueden usar para codificar datos o información de computadora en una pluralidad de identificadores, cada uno de los cuales puede representar uno o más bits de la información original. En algunos ejemplos, los métodos y sistemas de la presente divulgación codifican datos o información utilizando identificadores, cada uno de los cuales representa dos bits de la información original.

Los métodos anteriores para codificar información digital en ácidos nucleicos se han basado en la síntesis base por base de los ácidos nucleicos, lo que puede ser costoso y llevar mucho tiempo. Los métodos alternativos pueden mejorar la eficiencia, mejorar la viabilidad comercial del almacenamiento de información digital al reducir la dependencia de la síntesis de ácido nucleico base por base para codificar información digital y eliminar la síntesis de novo de distintas secuencias de ácidos nucleicos para cada nueva solicitud de almacenamiento de información.

Los nuevos métodos pueden codificar información digital (p. ej., código binario) en una pluralidad de identificadores o secuencias de ácidos nucleicos, que comprenden disposiciones combinatorias de componentes en lugar de depender en la síntesis de ácidos nucleicos base por base o de novo (p. ej., síntesis de fosforamidita). Como tales, las nuevas estrategias pueden producir un primer conjunto de secuencias (o componentes) de ácido nucleico distintas para la primera solicitud de almacenamiento de información y, posteriormente, pueden reutilizar las mismas secuencias (o componentes) de ácido nucleico para solicitudes de almacenamiento de información posteriores. Estos enfoques pueden reducir significativamente el costo del almacenamiento de información basado en ADN al reducir el papel de la síntesis de novo de secuencias de ácidos nucleicos en el proceso de codificación y escritura de información en ADN. Además, a diferencia de las implementaciones de síntesis base por base, como la elongación de ácidos nucleicos basada en la química de fosforamidita o la polimerasa sin molde, que puede utilizar la entrega cíclica de cada base a cada ácido nucleico en elongación, los nuevos métodos de escritura de información en ADN utilizando la construcción de identificadores a partir de componentes son procesos altamente paralelizables que no necesariamente utilizan la elongación cíclica del ácido nucleico. Por tanto, los nuevos métodos pueden aumentar la velocidad de

escritura de información digital en el ADN en comparación con los métodos más antiguos.

Métodos para codificar y escribir información en secuencia(s) de ácidos nucleicos

- 5 En un aspecto, la presente divulgación proporciona métodos para codificar información en secuencias de ácidos nucleicos. Un método para codificar información en secuencias de ácidos nucleicos puede comprender (a) traducir la información en una cadena de símbolos, (b) mapear la cadena de símbolos en una pluralidad de identificadores, y (c) construir una biblioteca de identificadores que comprende al menos un subconjunto de la pluralidad de identificadores. Un identificador individual de la pluralidad de identificadores puede comprender uno o más componentes. Un componente individual del uno o más componentes puede comprender una
- 10 secuencia de ácido nucleico. Cada símbolo en cada posición en la cadena de símbolos puede corresponder a un identificador distinto. El identificador individual puede corresponder a un símbolo individual en una posición individual en la cadena de símbolos. Además, un símbolo en cada posición de la cadena de símbolos puede corresponder a la ausencia de un identificador. Por ejemplo, en una cadena de símbolos binarios (p. ej., bits)
- 15 de '0' y '1', cada aparición de '0' puede corresponder a la ausencia de un identificador.

- En otro aspecto, la presente divulgación proporciona métodos para el almacenamiento de datos informáticos basado en ácidos nucleicos. Un método para el almacenamiento de datos informáticos basado en ácidos nucleicos puede comprender (a) recibir datos informáticos, (b) sintetizar moléculas de ácidos nucleicos que comprenden secuencias de ácidos nucleicos que codifican los datos informáticos, y (c) almacenar las
- 20 moléculas de ácidos nucleicos que tienen las secuencias de ácidos nucleicos. Los datos informáticos pueden codificarse en al menos un subconjunto de moléculas de ácidos nucleicos sintetizadas y no en una secuencia de cada una de las moléculas de ácidos nucleicos.

- 25 En otro aspecto, la presente divulgación proporciona métodos para escribir y almacenar información en secuencias de ácidos nucleicos. El método puede comprender, (a) recibir o codificar una biblioteca de identificadores virtuales que representa información, (b) construir físicamente la biblioteca de identificadores, y (c) almacenar una o más copias físicas de la biblioteca de identificadores en una o más localizaciones separadas. Un identificador individual de la biblioteca de identificadores puede comprender uno o más
- 30 componentes. Un componente individual del uno o más componentes puede comprender una secuencia de ácido nucleico.

- En otro aspecto, la presente divulgación proporciona métodos para el almacenamiento de datos informáticos basado en ácidos nucleicos. Un método para el almacenamiento de datos informáticos basado en ácidos nucleicos puede comprender (a) recibir datos informáticos, (b) sintetizar una molécula de ácido nucleico que comprende al menos una secuencia de ácido nucleico que codifica los datos informáticos, y (c) almacenar la
- 35 molécula de ácido nucleico que comprende la al menos una secuencia de ácido nucleico. La síntesis de la molécula de ácido nucleico puede realizarse en ausencia de síntesis de ácido nucleico base por base.

- 40 En otro aspecto, la presente divulgación proporciona métodos para escribir y almacenar información en secuencias de ácidos nucleicos. Un método para escribir y almacenar información en secuencias de ácidos nucleicos puede comprender, (a) recibir o codificar una biblioteca de identificadores virtuales que representa información, (b) construir físicamente la biblioteca de identificadores, y (c) almacenar una o más copias físicas de la biblioteca de identificadores en una o más localizaciones separadas. Un identificador individual de la
- 45 biblioteca de identificadores puede comprender uno o más componentes. Un componente individual del uno o más componentes puede comprender una secuencia de ácido nucleico.

- La FIG. 1 ilustra un proceso general para codificar información en secuencias de ácidos nucleicos, escribir información en las secuencias de ácidos nucleicos, leer información escrita en secuencias de ácidos nucleicos y descodificar la información leída. La información, o datos, digital se puede traducir en una o más cadenas de
- 50 símbolos. En un ejemplo, los símbolos son bits y cada bit puede tener un valor de '0' o '1'. Cada símbolo puede mapearse, o codificarse, a un objeto (p. ej., identificador) que representa ese símbolo. Cada símbolo puede estar representado por un identificador distinto. El identificador distinto puede ser una molécula de ácido nucleico formada por componentes. Los componentes pueden ser secuencias de ácidos nucleicos. La
- 55 información digital se puede escribir en secuencias de ácidos nucleicos generando una biblioteca de identificadores correspondiente a la información. La biblioteca de identificadores puede generarse físicamente construyendo físicamente los identificadores que corresponden a cada símbolo de la información digital. Se puede acceder a toda o cualquier parte de la información digital a la vez. En un ejemplo, se accede a un subconjunto de identificadores de una biblioteca de identificadores. El subconjunto de identificadores puede
- 60 leerse secuenciando e identificando los identificadores. Los identificadores identificados pueden asociarse con su símbolo correspondiente para descodificar los datos digitales.

- Un método para codificar y leer información usando el enfoque de la FIG. 1 puede, por ejemplo, incluir recibir una corriente de bits y mapear cada bit (bit con valor de bit de '1') en la corriente de bits a un identificador de
- 65 ácido nucleico distinto usando un rango de identificador o un índice de ácido nucleico. Construir un combinado de muestras de ácido nucleico, o biblioteca de identificadores, que comprende copias de los identificadores

que corresponden a valores de bits de 1 (y excluyendo identificadores para valores de bits de 0). La lectura de la muestra puede comprender el uso de métodos de biología molecular (p. ej., secuenciación, hibridación, PCR, etc.), determinar qué identificadores están representados en la biblioteca de identificadores y asignar valores de bits de '1' a los bits correspondientes a esos identificadores y valores de bits de '0' en otros lugares (refiriéndose nuevamente al rango de identificador para identificar los bits en la corriente de bits original al que corresponde cada identificador), descodificando así la información en la corriente de bits codificado original.

Al codificar una cadena de N bits distintos, se puede utilizar un número equivalente de secuencias de ácidos nucleicos únicas como posibles identificadores. Este enfoque para la codificación de información puede utilizar la síntesis de novo de identificadores (p. ej., moléculas de ácidos nucleicos) para cada nuevo elemento de información (cadena de N bits) a almacenar. En otros casos, el costo de sintetizar nuevos identificadores (equivalentes en número a o menos de N) para cada nuevo elemento de información a almacenar se puede reducir mediante la síntesis de novo por única vez y el mantenimiento posterior de todos los identificadores posibles, de modo que codificar nuevos elementos de información puede implicar seleccionar y mezclar mecánicamente entre sí identificadores presintetizados (o prefabricados) para formar una biblioteca de identificadores. En otros casos, tanto el costo de (1) la síntesis de novo de hasta N identificadores para cada nuevo elemento de información a almacenar o (2) el mantenimiento y selección de N posibles identificadores para cada nuevo elemento de información a almacenar, o cualquier combinación de los mismos, puede reducirse sintetizando y manteniendo un número (menor de N , y en algunos casos mucho menor de N) de secuencias de ácidos nucleicos y luego modificando estas secuencias mediante reacciones enzimáticas para generar hasta N identificadores para cada nuevo elemento de información a almacenar.

Los identificadores pueden diseñarse y seleccionarse racionalmente para facilitar las operaciones de lectura, escritura, acceso, copia y delección. Los identificadores pueden diseñarse y seleccionarse para minimizar errores de escritura, mutaciones, degradación y errores de lectura.

Las FIG. 2A y 2B ilustran esquemáticamente un método de ejemplo, denominado "datos en la dirección", de codificación de datos digitales en objetos o identificadores (p. ej., moléculas de ácidos nucleicos). La FIG. 2A ilustra la codificación de una corriente de bits en una biblioteca de identificadores en donde los identificadores individuales se construyen concatenando o ensamblando un único componente que especifica un rango de identificador con un único componente que especifica un valor de bit. En general, el método de datos en la dirección usa identificadores que codifican información de manera modular al comprender dos objetos: un objeto, el "objeto de valor de byte" (u "objeto de datos"), que identifica un valor de byte y un objeto, el "objeto de rango" (u "objeto de dirección"), que identifica el rango de identificador (o la posición relativa del byte en la corriente de bits original). La FIG. 2B ilustra un ejemplo del método de datos en la dirección en donde cada objeto de rango puede construirse combinatoriamente a partir de un conjunto de componentes y cada objeto de valor de byte puede construirse combinatoriamente a partir de un conjunto de componentes. Dicha construcción combinatoria de objetos de rango y valor de byte permite escribir más información en identificadores que si los objetos estuvieran formados solo por componentes únicos (p. ej., FIG. 2A).

Las FIG. 3A y 3B ilustran esquemáticamente otro método de ejemplo para codificar información digital en objetos o identificadores (p. ej., secuencias de ácidos nucleicos). La FIG. 3A ilustra la codificación de una corriente de bits en una biblioteca de identificadores en donde los identificadores se construyen a partir de componentes individuales que especifican el rango del identificador. La presencia de un identificador en un rango (o dirección) particular especifica un valor de bit de '1' y la ausencia de un identificador en un rango (o dirección) particular especifica un valor de bit de '0'. Este tipo de codificación puede usar identificadores que codifican únicamente el rango (la posición relativa de un bit en la corriente de bits original) y usar la presencia o ausencia de esos identificadores en una biblioteca de identificadores para codificar un valor de bit de '1' o '0', respectivamente. La lectura y descodificación de la información puede incluir identificar los identificadores presentes en la biblioteca de identificadores, asignar valores de bit de '1' a sus rangos correspondientes y asignar valores de bit de '0' en otros lugares. La FIG. 3B ilustra un método de codificación de ejemplo en el que cada identificador puede construirse combinatoriamente a partir de un conjunto de componentes de modo que cada construcción combinatoria posible especifique un rango. Dicha construcción combinatoria permite escribir más información en identificadores que si los identificadores estuvieran formados solo por componentes individuales (p. ej., FIG. 3A). Por ejemplo, un conjunto de componentes puede comprender cinco componentes distintos. Los cinco componentes distintos pueden ensamblarse para generar diez identificadores distintos, comprendiendo cada uno de ellos dos de los cinco componentes. Cada uno de los diez identificadores distintos puede tener un rango (o dirección) que corresponde a la posición de un bit en una corriente de bits. Una biblioteca de identificadores puede incluir el subconjunto de esos diez posibles identificadores que corresponde a las posiciones del valor de bit '1' y excluir el subconjunto de esos diez posibles identificadores que corresponde a las posiciones del valor de bit '0' dentro de una corriente de bits de longitud diez.

La FIG. 4 muestra un gráfico de contorno, en espacio log, de una relación entre el espacio combinatorio de posibles identificadores (C , eje x) y el número promedio de identificadores (k , eje y) que se construirán físicamente con el fin de almacenar información de un tamaño original dado en bits (D , líneas de contorno) usando el método de codificación mostrado en las FIG. 3A y 3B. Este gráfico asume que la información original

de tamaño D se recodifica en una cadena de C bits (donde C puede ser mayor que D) donde una cantidad de bits, k , tiene un valor de bit de '1'. Además, el gráfico asume que la codificación de información a ácido nucleico se realiza en la cadena de bits recodificada y que se construyen identificadores para posiciones donde el valor de bit es '1' y no se construyen identificadores para posiciones donde el valor de bit es '0'. Siguiendo estas asunciones, el espacio combinatorio de posibles identificadores tiene un tamaño C para identificar cada posición en la cadena de bits allí codificada, y el número de identificadores utilizados para codificar la cadena de bits de tamaño D es tal que $D = \log_2(C \text{ choose } k)$, donde $C \text{ choose } k$ puede ser la fórmula matemática para el número de formas de elegir k resultados desordenados de C posibilidades. Por tanto, a medida que el espacio combinatorio de posibles identificadores aumenta más allá del tamaño (en bits) de un determinado elemento de información, se puede utilizar un número decreciente de identificadores construidos físicamente para almacenar la información dada.

La FIG. 5 muestra un método general para escribir información en secuencias de ácidos nucleicos. Antes de escribir la información, la información puede traducirse en una cadena de símbolos y codificarse en una pluralidad de identificadores. La escritura de la información puede incluir configurar reacciones para producir posibles identificadores. Se puede configurar una reacción depositando entradas en un compartimento. Las entradas pueden comprender ácidos nucleicos, componentes, moldes, enzimas o reactivos químicos. El compartimento puede ser un pocillo, un tubo, una posición en una superficie, una cámara en un dispositivo de microfluidos o una gotita dentro de una emulsión. Se pueden configurar múltiples reacciones en múltiples compartimentos. Las reacciones pueden proceder para producir identificadores mediante incubación o ciclos a una temperatura programada. Las reacciones pueden eliminarse selectiva o ubicuamente (p. ej., delecionarse). Las reacciones también pueden interrumpirse, consolidarse y purificarse de forma selectiva o ubicua para recopilar sus identificadores en un combinado. Se pueden recopilar identificadores de múltiples bibliotecas de identificadores en el mismo combinado. Un identificador individual puede incluir un código de barras o una etiqueta para identificar a qué biblioteca de identificadores pertenece. Alternativamente o, además, el código de barras puede incluir metadatos para la información codificada. También se pueden incluir identificadores o ácidos nucleicos suplementarios en un combinado de identificadores junto con una biblioteca de identificadores. Los ácidos nucleicos o identificadores suplementarios pueden incluir metadatos para la información codificada o servir para ofuscar u ocultar la información codificada.

Un rango de identificador (p. ej., índice de ácido nucleico) puede comprender un método o clave para determinar el orden de los identificadores. El método puede comprender una tabla de consulta con todos los identificadores y su rango correspondiente. El método también puede comprender una tabla de consulta con el rango de todos los componentes que constituyen identificadores y una función para determinar el orden de cualquier identificador que comprenda una combinación de esos componentes. Dicho método puede denominarse orden lexicográfico y puede ser análogo a la manera en que se ordenan alfabéticamente las palabras de un diccionario. En el método de codificación de datos en la dirección, el rango del identificador (codificado por el objeto de rango del identificador) se puede usar para determinar la posición de un byte (codificado por el objeto de valor de byte del identificador) dentro de una corriente de bits. En un método alternativo, el rango de identificador (codificado por el propio identificador completo) para un identificador presente puede usarse para determinar la posición del valor de bit de '1' dentro de una corriente de bits.

Una clave puede asignar bits distintos a subconjuntos únicos de identificadores (p. ej., moléculas de ácidos nucleicos) dentro de una muestra. Por ejemplo, en una forma simple, una clave puede asignar cada bit de un byte a una secuencia de ácido nucleico única que especifica la posición del bit, y luego la presencia o ausencia de esa secuencia de ácido nucleico dentro de una muestra puede especificar el valor de bit de 1 o 0, respectivamente. La lectura de la información codificada de la muestra de ácido nucleico puede comprender cualquier número de técnicas de biología molecular, incluidas la secuenciación, la hibridación o la PCR. En algunas realizaciones, la lectura del conjunto de datos codificados puede comprender reconstruir una parte del conjunto de datos o reconstruir el conjunto de datos codificados completo a partir de cada muestra de ácido nucleico. Cuando la secuencia puede leerse, el índice de ácido nucleico puede usarse junto con la presencia o ausencia de una secuencia de ácido nucleico única y la muestra de ácido nucleico puede descodificarse en una corriente de bits (p. ej., cada cadena de bits, byte, bytes o cadena de bytes).

Los identificadores pueden construirse ensamblando combinatoriamente secuencias de ácidos nucleicos componentes. Por ejemplo, la información puede codificarse tomando un conjunto de moléculas de ácidos nucleicos (p. ej., identificadores) de un grupo definido de moléculas (p. ej., espacio combinatorio). Cada posible identificador del grupo definido de moléculas puede ser un ensamblaje de secuencias de ácidos nucleicos (p. ej., componentes) de un conjunto prefabricado de componentes que pueden dividirse en capas. Cada identificador individual puede construirse concatenando un componente de cada capa en un orden fijo. Por ejemplo, si hay M capas y cada capa puede tener n componentes, entonces se pueden construir hasta $C = n^M$ identificadores únicos y se pueden codificar y almacenar hasta 2^C diferentes elementos de información, o C bits. Por ejemplo, el almacenamiento de un megabit de información puede utilizar 1×10^6 identificadores distintos o un espacio combinatorio de tamaño $C = 1 \times 10^6$. Los identificadores de este ejemplo se pueden ensamblar a partir de una variedad de componentes organizados de diferentes maneras. Los ensamblajes pueden realizarse a partir de $M = 2$ capas prefabricadas, conteniendo cada una $n = 1 \times 10^3$ componentes.

Alternativamente, se pueden hacer ensamblajes a partir de $M = 3$ capas, conteniendo cada una $n = 1 \times 10^2$ componentes. Como ilustra este ejemplo, la codificación de la misma cantidad de información usando un mayor número de capas puede permitir que el número total de componentes sea menor. El uso de un número menor de componentes totales puede resultar ventajoso en términos de costo de escritura.

5

En un ejemplo, se puede comenzar con dos conjuntos de secuencias o capas de ácido nucleico únicas, X e Y, cada una con componentes x e y (p. ej., secuencias de ácidos nucleicos), respectivamente. Cada secuencia de ácido nucleico de X puede ensamblarse en cada secuencia de ácido nucleico de Y. Aunque el número total de secuencias de ácidos nucleicos mantenido en los dos conjuntos puede ser la suma de x e y , el número total de moléculas de ácidos nucleicos y, por tanto, posibles identificadores, que se pueden generar puede ser el producto de x e y . Se pueden generar incluso más secuencias de ácidos nucleicos (p. ej., identificadores) si las secuencias de X se pueden ensamblar con las secuencias de Y en cualquier orden. Por ejemplo, el número de secuencias de ácidos nucleicos (p. ej., identificadores) generadas puede ser el doble del producto de x e y si el orden de ensamblaje es programable. Este conjunto de todas las posibles secuencias de ácidos nucleicos que pueden generarse puede denominarse XY. El orden de las unidades ensambladas de secuencias de ácidos nucleicos únicas en XY se puede controlar usando ácidos nucleicos con extremos 5' y 3' distintos, y se puede realizar digestión de restricción, ligación, reacción en cadena de la polimerasa (PCR) y secuenciación con respecto a los distintos extremos 5' y 3' de las secuencias. Dicho enfoque puede reducir el número total de secuencias de ácidos nucleicos (p. ej., componentes) utilizadas para codificar N bits distintos, codificando información en las combinaciones y órdenes de sus productos de ensamblaje. Por ejemplo, para codificar 100 bits de información, se pueden ensamblar dos capas de 10 moléculas de ácidos nucleicos distintas (p. ej., componente) en un orden fijo para producir 10×10 o 100 moléculas de ácidos nucleicos distintas (p. ej., identificadores), o una capa de 5 moléculas de ácidos nucleicos distintas (p. ej., componentes) y otra capa de 10 moléculas de ácidos nucleicos distintas (p. ej., componentes) se pueden ensamblar en cualquier orden para producir 100 moléculas de ácidos nucleicos distintas (p. ej., identificadores).

25

Las secuencias de ácidos nucleicos (p. ej., componentes) dentro de cada capa pueden comprender una secuencia única (o distinta), o código de barras, en el medio, una región de hibridación común en un extremo y otra región de hibridación común en el otro extremo. El código de barras puede contener un número suficiente de nucleótidos para identificar de forma única cada secuencia dentro de la capa. Por ejemplo, típicamente hay cuatro nucleótidos posibles para cada posición de base dentro de un código de barras. Por lo tanto, un código de barras de tres bases puede identificar de forma única $4^3 = 64$ secuencias de ácidos nucleicos. Los códigos de barras pueden diseñarse para generarse aleatoriamente. Alternativamente, los códigos de barras pueden diseñarse para evitar secuencias que puedan crear complicaciones en la química de construcción de identificadores o secuenciación. Además, los códigos de barras pueden diseñarse de modo que cada uno pueda tener una distancia de Hamming mínima de los otros códigos de barras, disminuyendo así la probabilidad de que mutaciones de resolución base o errores de lectura puedan interferir con la identificación adecuada del código de barras.

30

35

La región de hibridación en un extremo de la secuencia de ácido nucleico (p. ej., componente) puede ser diferente en cada capa, pero la región de hibridación puede ser la misma para cada miembro dentro de una capa. Las capas adyacentes son aquellas que tienen regiones de hibridación complementarias en sus componentes que les permiten interactuar entre sí. Por ejemplo, cualquier componente de la capa X puede unirse a cualquier componente de la capa Y porque pueden tener regiones de hibridación complementarias. La región de hibridación en el extremo opuesto puede tener el mismo propósito que la región de hibridación en el primer extremo. Por ejemplo, cualquier componente de la capa Y puede unirse a cualquier componente de la capa X en un extremo y a cualquier componente de la capa Z en el extremo opuesto.

40

45

Las FIG. 6A y 6B ilustran un método de ejemplo, denominado "esquema de producto", para construir identificadores (p. ej., moléculas de ácidos nucleicos) ensamblando combinatoriamente un componente distinto (p. ej., secuencia de ácido nucleico) de cada capa en un orden fijo. La FIG. 6A ilustra la arquitectura de identificadores construidos usando el esquema de producto. Se puede construir un identificador combinando un único componente de cada capa en un orden fijo. Para M capas, cada una con N componentes, hay N^M posibles identificadores. La FIG. 6B ilustra un ejemplo del espacio combinatorio de identificadores que se puede construir usando el esquema de producto. En un ejemplo, se puede generar un espacio combinatorio a partir de tres capas, comprendiendo cada una tres componentes distintos. Los componentes se pueden combinar de manera que un componente de cada capa se pueda combinar en un orden fijo. Todo el espacio combinatorio para este método de ensamblaje puede comprender veintisiete identificadores posibles.

50

55

Las FIG. 7-10 ilustran métodos químicos para implementar el esquema del producto (véase la FIG. 6). Se pueden usar los métodos representados en las FIG. 7-10, junto con cualesquiera otros métodos para ensamblar dos o más componentes distintos en un orden fijo, por ejemplo, para producir uno cualquiera o más identificadores en una biblioteca de identificadores. Los identificadores se pueden construir utilizando cualesquiera de los métodos de implementación descritos en las FIG. 7-10, en cualquier momento durante los métodos o sistemas divulgados en la presente memoria. En algunos casos, todo o una parte del espacio combinatorio de posibles identificadores puede construirse antes de codificar o escribir la información digital, y

60

65

luego el proceso de escritura puede implicar seleccionar y combinar mecánicamente los identificadores (que codifican la información) a partir del conjunto ya existente. En otros casos, los identificadores pueden construirse después de que hayan ocurrido una o más etapas del proceso de codificación o escritura de datos (es decir, mientras se escribe la información).

5

Se pueden usar reacciones enzimáticas para ensamblar componentes de las diferentes capas o conjuntos. El ensamblaje puede ocurrir en una reacción de un solo recipiente porque los componentes (p. ej., secuencias de ácidos nucleicos) de cada capa tienen regiones de hibridación o unión específicas para componentes de capas adyacentes. Por ejemplo, una secuencia de ácido nucleico (p. ej., componente) X1 de la capa X, una secuencia de ácido nucleico Y1 de la capa Y y una secuencia de ácido nucleico Z1 de la capa Z pueden formar la molécula de ácido nucleico ensamblada (p. ej., identificador) X1Y1Z1. Además, se pueden ensamblar múltiples moléculas de ácidos nucleicos (p. ej., identificadores) en una reacción incluyendo múltiples secuencias de ácidos nucleicos de cada capa. Por ejemplo, incluir Y1 e Y2 en la reacción de un solo recipiente del ejemplo anterior puede producir dos productos ensamblados (p. ej., identificadores), X1Y1Z1 y X1Y2Z1. Esta multiplexación de reacción se puede utilizar para acelerar el tiempo de escritura para la pluralidad de identificadores que se construyen físicamente. El ensamblaje de las secuencias de ácidos nucleicos se puede realizar en un período de tiempo que es menor o igual a aproximadamente 1 día, 12 horas, 10 horas, 9 horas, 8 horas, 7 horas, 6 horas, 5 horas, 4 horas, 3 horas, 2 horas o 1 hora. La precisión de los datos codificados puede ser al menos aproximadamente o igual a aproximadamente el 90 %, 95 %, 96 %, 97 %, 98 %, 99 % o mayor.

20

Los identificadores se pueden construir de acuerdo con el esquema del producto usando la reacción en cadena de la polimerasa de extensión superpuesta (OEPCR), como se ilustra en la FIG. 7. Cada componente en cada capa puede comprender una secuencia de ácido nucleico bicatenario o monocatenario (como se representa en la figura) con una región de hibridación común en el extremo de la secuencia que puede ser homóloga y/o complementaria a la región de hibridación común en el extremo de la secuencia de componentes de una capa adyacente. Se puede construir un identificador individual concatenando un componente (p. ej., secuencia única) de una capa X (o capa 1) que comprende los componentes X₁ - X_A, un segundo componente (p. ej., secuencia única) de una capa Y (o capa 2) que comprende Y₁ - Y_A, y un tercer componente (p. ej., secuencia única) de la capa Z (o capa 3) que comprende Z₁ - Z_B. Los componentes de la capa X pueden tener un extremo 3' que comparte complementariedad con el extremo 3' de los componentes de la capa Y. Por lo tanto, los componentes monocatenarios de las capas X e Y pueden hibridarse entre sí en el extremo 3' y pueden extenderse usando PCR para generar una molécula de ácido nucleico bicatenaria. La molécula de ácido nucleico bicatenaria generada se puede fundir para generar un extremo 3' que comparte complementariedad con un extremo 3' de un componente de la capa Z. Un componente de la capa Z se puede hibridar con la molécula de ácido nucleico generada y se puede extender para generar un identificador único que comprende un único componente de las capas X, Y y Z en un orden fijo. Se puede implementar la selección del tamaño del ADN (p. ej., con extracción en gel) o la reacción en cadena de la polimerasa (PCR) con cebadores que flanquean las capas más externas para aislar los productos identificadores de otros subproductos que pueden formarse en la reacción.

40

Los identificadores se pueden ensamblar de acuerdo con el esquema del producto usando ligación de extremos cohesivos, como se ilustra en la FIG. 8. Se pueden utilizar tres capas, comprendiendo cada una, componentes bicatenarios (p. ej., ADN bicatenario (ADNs)) con protuberancias 3' monocatenarias, para ensamblar identificadores distintos. Por ejemplo, identificadores que comprenden un componente de la capa X (o capa 1) que comprende los componentes X₁ - X_A, un segundo componente de la capa Y (o capa 2) que comprende Y₁ - Y_B, y un tercer componente de la capa Z (o capa 3) que comprende Z₁ - Z_C. Para combinar componentes de la capa X con componentes de la capa Y, los componentes de la capa X pueden comprender una protuberancia 3' común, FIG. 8 etiquetado como **a**, y los componentes en la capa Y pueden comprender una protuberancia 3' común y complementaria, **a***. Para combinar componentes de la capa Y con componentes de la capa Z, los elementos de la capa Y pueden comprender una protuberancia 3' común, FIG. 8 etiquetado como **b**, y los elementos en la capa Z pueden comprender una protuberancia 3' común y complementaria, **b***. La protuberancia 3' en los componentes de la capa X puede ser complementaria al extremo 3' de los componentes de la capa Y y la otra protuberancia 3' en los componentes de la capa Y puede ser complementaria al extremo 3' de los componentes de la capa Z, permitiendo que los componentes se hibriden y se ligan. Como tales, los componentes de la capa X no pueden hibridarse con otros componentes de la capa X o la capa Z, y de manera similar, los componentes de la capa Y no pueden hibridarse con otros elementos de la capa Y. Además, un único componente de la capa Y puede ligarse a un único componente de la capa X y un único componente de la capa Z, asegurando la formación de un identificador completo. Se puede implementar la selección del tamaño del ADN (p. ej., con extracción en gel) o PCR con cebadores que flanquean las capas más externas para aislar los productos identificadores de otros subproductos que pueden formarse en la reacción.

60

Los extremos cohesivos para la ligación de extremos cohesivos se pueden generar tratando los componentes de cada capa con endonucleasas de restricción. En algunas realizaciones, los componentes de múltiples capas pueden generarse a partir de un conjunto "parental" de componentes. Por ejemplo, una realización en donde un único conjunto parental de componentes bicatenarios puede tener sitios de restricción complementarios en

65

- cada extremo (p. ej., sitios de restricción para BamHI y BglII). Se pueden seleccionar dos componentes cualesquiera para el ensamblaje y se pueden digerir individualmente con una u otra enzima de restricción complementaria (p. ej., BglII o BamHI) dando lugar a extremos cohesivos complementarios que se pueden ligar entre sí dando como resultado una cicatriz inerte. La secuencia de ácido nucleico producto puede comprender
- 5 los sitios de restricción complementarios en cada extremo (p. ej., BamHI en el extremo 5' y BglII en el extremo 3'), y pueden ligarse además a otro componente del conjunto parental siguiendo el mismo proceso. Este proceso puede tener un ciclo indefinido. Si el parental comprende *N* componentes, entonces cada ciclo puede ser equivalente a añadir una capa adicional de *N* componentes al esquema del producto.
- 10 Un método para usar ligación para construir una secuencia de ácidos nucleicos que comprende elementos del conjunto X (p. ej., conjunto 1 de ADNds) y elementos del conjunto Y (p. ej., conjunto 2 de ADNds) puede comprender las etapas de obtener o construir dos o más combinados (p. ej., conjunto 1 de ADNds y conjunto 2 de ADNds) de secuencias bicatenarias en donde un primer conjunto (p. ej., conjunto 1 de ADNds) comprende
- 15 un extremo cohesivo (p. ej., *a*) y un segundo conjunto (p. ej., conjunto 2 de ADNds) comprende un extremo cohesivo (p. ej., *a*^{*}) que es complementario al extremo cohesivo del primer conjunto. Cualquier ADN del primer conjunto (p. ej., conjunto 1 de ADNds) y cualquier subconjunto de ADN del segundo conjunto (p. ej., conjunto 2 de ADNds) se pueden combinar y ensamblar y luego ligar entre sí para formar un único ADN bicatenario con un elemento del primer conjunto y un elemento del segundo conjunto.
- 20 Los identificadores se pueden ensamblar de acuerdo con el esquema del producto usando recombinación específica de sitio, como se ilustra en la FIG. 9. Los identificadores se pueden construir ensamblando componentes de tres capas diferentes. Los componentes de la capa X (o capa 1) pueden comprender moléculas bicatenarias con un sitio de recombinasa attB_x en un lado de la molécula, los componentes de la capa Y (o capa 2) pueden comprender moléculas bicatenarias con un sitio de recombinasa attP_x en un lado y
- 25 un sitio de recombinasa attB_y en el otro lado, y los componentes en la capa Z (o capa 3) pueden comprender un sitio de recombinasa attP_y en un lado de la molécula. Los sitios attB y attP dentro de un par, como lo indican sus subíndices, son capaces de recombinarse en presencia de su correspondiente enzima recombinasa. Se puede combinar un componente de cada capa de manera que un componente de la capa X se asocie con un componente de la capa Y, y un componente de la capa Y se asocie con un componente de la capa Z. La aplicación de una o más enzimas recombinasas puede recombinar los componentes para generar un
- 30 identificador bicatenario que comprende los componentes ordenados. Se puede implementar la selección del tamaño del ADN (p. ej., con extracción en gel) o PCR con cebadores que flanquean las capas más externas para aislar los productos identificadores de otros subproductos que pueden formarse en la reacción. En general, se pueden usar múltiples pares attB y attP ortogonales, y cada par se puede usar para ensamblar un
- 35 componente a partir de una capa adicional. Para la familia de recombinasas de serina grande, se pueden generar hasta seis pares attB y attP ortogonales por recombinasas, y también se pueden implementar múltiples recombinasas ortogonales. Por ejemplo, se pueden ensamblar trece capas utilizando doce pares attB y attP ortogonales, seis pares ortogonales de cada una de dos grandes recombinasas de serina, tales como BxbI y PhiC31. La ortogonalidad de los pares attB y attP garantiza que un sitio attB de un par no reaccione con un
- 40 sitio attP de otro par. Esto permite ensamblar componentes de diferentes capas en un orden fijo. Las reacciones de recombinación mediadas por recombinasa pueden ser reversibles o irreversibles dependiendo del sistema de recombinasa implementado. Por ejemplo, la gran familia de recombinasas de serina cataliza reacciones de recombinación irreversibles sin requerir cofactores de alta energía, mientras que la familia de recombinasas de tirosina cataliza reacciones reversibles.
- 45 Se pueden construir identificadores de acuerdo con el esquema del producto usando ligación dirigida por molde (TDL), como se muestra en la FIG. 10A. La ligación dirigida por molde utiliza secuencias de ácido nucleico monocatenario, denominadas "molde" o "grapas", para facilitar la ligación ordenada de componentes para formar identificadores. Los moldes se hibridan simultáneamente con componentes de capas adyacentes y los
- 50 mantienen adyacentes entre sí (extremo 3' contra extremo 5') mientras una ligasa los liga. En el ejemplo de la FIG. 10A, se combinan tres capas o conjuntos de componentes monocatenarios. Una primera capa de componentes (p. ej., capa X o capa 1) que comparten secuencias comunes *a* en su extremo 3', que son complementarias a las secuencias *a*^{*}; una segunda capa de componentes (p. ej., capa Y o capa 2) que comparten secuencias comunes *b* y *c* en sus extremos 5' y 3', respectivamente, que son complementarias a
- 55 las secuencias *b*^{*} y *c*^{*}; una tercera capa de componentes (p. ej., capa Z o capa 3) que comparten una secuencia común *d* en su extremo 5', que puede ser complementaria a las secuencias *d*^{*}; y un conjunto de dos moldes o "grapas" donde la primera grapa comprende la secuencia *a*^{*}*b*^{*} (5' a 3') y la segunda grapa comprende una secuencia *c*^{*}*d*^{*} (5' a 3'). En este ejemplo, se pueden seleccionar uno o más componentes de cada capa y mezclarlos en una reacción con las grapas, lo que, mediante hibridación complementaria, puede facilitar la
- 60 ligación de un componente de cada capa en un orden definido para formar un identificador. Se puede implementar la selección del tamaño del ADN (p. ej., con extracción en gel) o PCR con cebadores que flanquean las capas más externas para aislar los productos identificadores de otros subproductos que pueden formarse en la reacción.
- 65 La FIG. 10B muestra un histograma de los números de copias (abundancias) de 256 secuencias de ácidos nucleicos distintas, cada una de las cuales se ensambló con TDL de 6 capas. Las capas del borde (capas

primera y final) tenían cada una un componente, y cada una de las capas internas (las cuatro capas restantes) tenía cuatro componentes. Cada componente de la capa del borde tenía 28 bases, incluida una región de hibridación de 10 bases. Cada componente de la capa interna tenía 30 bases, incluida una región de hibridación común de 10 bases en el extremo 5', una región variable (código de barras) de 10 bases y una región de hibridación común de 10 bases en el extremo 3'. Cada una de las tres cadenas molde tenía una longitud de 20 bases. Las 256 secuencias distintas se ensamblaron todas de forma múltiple con una reacción que contenía todos los componentes y moldes, polinucleótido quinasa T4 (para fosforilar los componentes) y ligasa T4, ATP y otros reactivos de reacción adecuados. La reacción se incubó a 37 grados durante 30 minutos y luego a temperatura ambiente durante 1 hora. Se añadieron adaptadores de secuenciación al producto de reacción con PCR y el producto se secuenció con un instrumento Illumina MiSeq. Se muestra el número de copias relativo de cada secuencia ensamblada distinta del total de 192.910 lecturas de secuencias ensambladas. Otras realizaciones de este método pueden usar componentes bicatenarios, donde los componentes se funden inicialmente para formar versiones monocatenarias que pueden hibridar con las grapas. Se pueden usar otras realizaciones o derivados de este método (es decir, TDL) para construir un espacio combinatorio de identificadores más complejo que el que se puede lograr en el esquema del producto.

Los identificadores se pueden construir de acuerdo con el esquema del producto utilizando otras implementaciones químicas diversas, incluido el ensamblaje golden gate, ensamblaje de Gibson y ensamblaje de reacción de ciclación de ligasa.

Las FIG. 11A y 11B ilustran esquemáticamente un método de ejemplo, denominado "esquema de permutación", para construir identificadores (p. ej., moléculas de ácidos nucleicos) con componentes permutados (p. ej., secuencias de ácidos nucleicos). La FIG. 11A ilustra la arquitectura de identificadores construidos usando el esquema de permutación. Se puede construir un identificador combinando un único componente de cada capa en un orden programable. La FIG. 11B ilustra un ejemplo del espacio combinatorio de identificadores que se puede construir usando el esquema de permutación. En un ejemplo, se puede generar un espacio combinatorio de tamaño seis a partir de tres capas, comprendiendo cada una un componente distinto. Los componentes pueden concatenarse en cualquier orden. En general, con M capas, cada una con N componentes, el esquema de permutación permite un espacio combinatorio de $N^M M!$ identificadores totales.

La FIG. 11C ilustra una implementación de ejemplo del esquema de permutación con ligación dirigida por molde (TDL). Los componentes de múltiples capas se ensamblan entre los componentes fijos del extremo izquierdo y del derecho, lo que se conoce como soportes de borde. Estos soportes de borde son los mismos para todos los identificadores en el espacio combinatorio y, por lo tanto, pueden añadirse como parte de la mezcla maestra de reacción para la implementación. Existen moldes o grapas para cualquier posible unión entre dos capas o soportes cualesquiera, de modo que el orden en el que los componentes de diferentes capas se incorporan en un identificador en la reacción depende de los moldes seleccionados para la reacción. Con el fin de permitir cualquier posible permutación de capas para M capas, puede haber $M^2 + 2M$ grapas seleccionables distintas para cada unión posible (incluidas las uniones con los soportes). M de esos moldes (sombreados en gris) forman uniones entre las capas y ellas mismas y pueden excluirse a los efectos del ensamblaje de permutación como se describe en la presente memoria. Sin embargo, su inclusión puede permitir un espacio combinatorio más grande con identificadores que comprenden componentes repetidos como se ilustra en las FIG. 11D-G. Se puede implementar la selección del tamaño del ADN (por ejemplo, con extracción en gel) o PCR con cebadores dirigidos a los soportes de borde para aislar los productos identificadores de otros subproductos que pueden formarse en la reacción.

Las FIG. 11D-G ilustran métodos de ejemplo de cómo se puede expandir el esquema de permutación para incluir ciertos casos de identificadores con componentes repetidos. La FIG. 11D muestra un ejemplo de cómo la forma de implementación FIG. 11C puede usarse para construir identificadores con componentes permutados y repetidos. Por ejemplo, un identificador puede comprender tres componentes en total ensamblados a partir de dos componentes distintos. En este ejemplo, un componente de una capa puede estar presente múltiples veces en un identificador. Se pueden lograr concatenaciones adyacentes del mismo componente usando una grapa con regiones de hibridación complementarias adyacentes tanto para el extremo 3' como para el extremo 5' del mismo componente, tal como la grapa a*b* (5' a 3') en la figura. En general, para M capas, existen M grapas de este tipo. La incorporación de componentes repetidos con esta implementación puede generar secuencias de ácidos nucleicos de más de una longitud (es decir, que comprenden uno, dos, tres, cuatro o más componentes) que se ensamblan entre los soportes de borde, como se demuestra en la FIG. 11E. La FIG. 11E muestra cómo la implementación de ejemplo de la FIG. 11D puede conducir a secuencias de ácidos nucleicos no diana, además del identificador, que se ensamblan entre los soportes de borde. El identificador apropiado no se puede aislar de una secuencia de ácido nucleico no diana mediante PCR porque comparten los mismos sitios de unión del cebador en el borde. Sin embargo, en este ejemplo, se puede implementar la selección del tamaño del ADN (p. ej., con extracción en gel) para aislar el identificador diana (p. ej., la segunda secuencia desde arriba) de las secuencias no diana, ya que cada secuencia de ácido nucleico ensamblada se puede diseñar para tener una longitud única (p. ej., si todos los componentes tienen la misma longitud). La FIG. 11F muestra otro ejemplo en el que la construcción de un identificador con componentes repetidos puede generar múltiples secuencias de ácidos nucleicos con secuencias de bordes

iguales, pero longitudes distintas en la misma reacción. En este método, se pueden usar moldes que ensamblan componentes en una capa con componentes en otras capas en un patrón alterno. Como con el método mostrado en la FIG. 11E, la selección de tamaño se puede usar para seleccionar identificadores de la longitud diseñada. La FIG. 11G muestra un ejemplo en el que la construcción de un identificador con componentes repetidos puede generar múltiples secuencias de ácidos nucleicos con secuencias de borde iguales y para algunas secuencias de ácidos nucleicos (p. ej., la tercera y cuarta desde arriba y la sexta y séptima desde arriba), longitudes iguales. En este ejemplo, aquellas secuencias de ácidos nucleicos que comparten longitudes iguales pueden excluirse de que ambos sean identificadores individuales, ya que puede no ser posible construir una sin construir también la otra, incluso si se implementa la PCR y la selección del tamaño del ADN.

Las FIG. 12A - 12D ilustran esquemáticamente un método de ejemplo, denominado "esquema MChooseK", para construir identificadores (p. ej., moléculas de ácidos nucleicos) con cualquier número, K , de componentes ensamblados (p. ej., secuencias de ácidos nucleicos) de un número mayor, M , de posibles componentes. La FIG. 12A ilustra la arquitectura de identificadores construidos usando el esquema de MchooseK. Usando este método, los identificadores se construyen ensamblando un componente de cada capa en cualquier subconjunto de todas las capas (p. ej., se eligen componentes de k capas de M capas posibles). La FIG. 12B ilustra un ejemplo del espacio combinatorio de identificadores que se puede construir usando el esquema de MchooseK. En este esquema de ensamblaje, el espacio combinatorio puede comprender N^K MchooseK posibles identificadores para M capas, N componentes por capa y una longitud de identificador de K componentes. En un ejemplo, si hay cinco capas, comprendiendo cada una un componente, entonces se pueden ensamblar hasta diez identificadores distintos que comprenden dos componentes cada uno.

El esquema MchooseK se puede implementar usando ligación dirigida por molde, como se muestra en la FIG. 12C. Al igual que con la implementación de TDL para el esquema de permutación (FIG. 11C), los componentes en este ejemplo se ensamblan entre soportes de borde que pueden incluirse o no en la mezcla maestra de reacción. Los componentes se pueden dividir en M capas, por ejemplo, $M = 4$ capas con rango predefinido de 2 a M , donde el soporte del borde izquierdo puede tener el rango 1 y el soporte del borde derecho puede tener el rango $M+1$. Los moldes comprenden secuencias de ácidos nucleicos para la ligación 3' a 5' de dos componentes cualesquiera con un rango inferior a un rango superior, respectivamente. Hay $((M+1)^2 + M+1)/2$ de dichos moldes. Se puede construir un identificador individual de cualquier componente K de distintas capas combinando esos componentes seleccionados en una reacción de ligación con las grapas $K+1$ correspondientes utilizadas para unir los componentes K junto con los soportes de borde en su orden de rango. Dicha configuración de reacción puede producir la secuencia de ácido nucleico correspondiente al identificador diana entre los soportes de borde. Alternativamente, se puede combinar una mezcla de reacción que comprende todos los moldes con los componentes seleccionados para ensamblar el identificador diana. Este método alternativo puede generar varias secuencias de ácidos nucleicos con las mismas secuencias de borde, pero longitudes distintas (si todas las longitudes de los componentes son iguales), como se ilustra en la FIG. 12D. El identificador diana (abajo) se puede aislar a partir de secuencias de ácidos nucleicos subproductos por tamaño.

Las FIG. 13A y 13B ilustran esquemáticamente un método de ejemplo, denominado "esquema de partición" para construir identificadores con componentes divididos. La FIG. 13A muestra un ejemplo del espacio combinatorio de identificadores que se puede construir usando el esquema de partición. Se puede construir un identificador individual ensamblando un componente de cada capa en un orden fijo con la colocación opcional de cualquier partición (componente especialmente clasificado) entre dos componentes cualesquiera de capas diferentes. Por ejemplo, un conjunto de componentes se puede organizar en un componente de partición y cuatro capas que contienen un componente cada una. Se puede combinar un componente de cada capa en un orden fijo y se puede ensamblar un único componente de partición en varias localizaciones entre las capas. Un identificador en este espacio combinatorio puede no comprender componentes de partición, un componente de partición entre los componentes de la primera y segunda capa, una partición entre los componentes de la segunda y tercera capa, y así sucesivamente para formar un espacio combinatorio de ocho posibles identificadores. En general, con M capas, cada una con N componentes y p componentes de partición, hay $N^K(p+1)^{M-1}$ posibles identificadores que se pueden construir. Este método puede generar identificadores de varias longitudes.

La FIG. 13B muestra una implementación de ejemplo del esquema de partición usando ligación dirigida por molde. Los moldes comprenden secuencias de ácidos nucleicos para ligar un componente de cada una de las M capas en un orden fijo. Para cada componente de partición, existen pares adicionales de moldes que permiten que el componente de partición se ligue entre los componentes de dos capas adyacentes cualesquiera. Por ejemplo, un par de moldes de modo que un molde (con secuencia g^*b^* (5' a 3'), por ejemplo) en un par permite que el extremo 3' de la capa 1 (con secuencia b) se ligue al extremo 5' del componente de partición (con secuencia g) y de manera que el segundo molde en el par (con secuencia c^*h^* (5' a 3'), por ejemplo) permita que el extremo 3' del componente de partición (con secuencia h) se ligue al extremo 5' de la capa 2 (con secuencia c). Para insertar una partición entre dos componentes cualesquiera de capas adyacentes, el molde estándar para ligar esas capas entre sí se puede excluir en la reacción y el par de moldes

para ligar la partición en esa posición se puede seleccionar en la reacción. En el ejemplo actual, la toma como diana del componente de partición entre la capa 1 y la capa 2 puede usar el par de moldes c^*h^* (5' a 3') y g^*b^* (5' a 3') para seleccionar la reacción en lugar del molde c^*b^* (5' a 3'). Los componentes se pueden ensamblar entre soportes de borde que se pueden incluir en la mezcla de reacción (junto con sus moldes correspondientes para ligar a la primera y a la capa M , respectivamente). En general, se puede utilizar un total de alrededor de $M-1+2*p*(M-1)$ moldes seleccionables para este método para M capas y p componentes de partición. Esta implementación del esquema de partición puede generar varias secuencias de ácidos nucleicos en una reacción con las mismas secuencias de borde, pero de distintas longitudes. El identificador diana se puede aislar a partir de secuencias de ácidos nucleicos subproductos mediante la selección del tamaño del ADN. Específicamente, puede haber exactamente un producto de secuencia de ácido nucleico con exactamente M componentes de capa. Si los componentes de la capa se diseñan lo suficientemente grandes en comparación con los componentes de la partición, puede ser posible definir una región de selección de tamaño universal mediante la cual el identificador (y ninguno de los subproductos no diana) puede seleccionarse independientemente de la partición particular de los componentes dentro del identificador, permitiendo así aislar múltiples identificadores particionados de múltiples reacciones en la misma etapa de selección de tamaño.

Las FIG. 14A y 14B ilustran esquemáticamente un método de ejemplo, denominado "esquema de cadena sin restricciones" o "USS", para construir identificadores formados por cualquier cadena de componentes a partir de un número de componentes posibles. La FIG. 14A muestra un ejemplo del espacio combinatorio de identificadores de longitud de 3 componentes (o 4 soportes) que pueden construirse usando el esquema de cadena sin restricciones. El esquema de cadena sin restricciones construye un identificador individual de longitud K componentes con uno o más componentes distintos, cada uno tomado de una o más capas, donde cada componente distinto puede aparecer en cualquiera de las posiciones de los K componentes en el identificador (permitiendo repeticiones). Por ejemplo, para dos capas, comprendiendo cada una un componente, hay ocho posibles identificadores de longitud de 3 componentes. En general, con M capas, cada una con un componente, hay M^K posibles identificadores de longitud K componentes. La FIG. 14B muestra una implementación de ejemplo del esquema de cadena sin restricciones usando ligación dirigida por molde. En este método, están presentes en la mezcla de reacción $K+1$ componentes de ADN monocatenario y de soporte ordenado (incluidos dos soportes de borde y $K-1$ soportes internos). Un identificador individual comprende un único componente ligado entre cada par de soportes adyacentes. Por ejemplo, un componente ligado entre los soportes A y B, un componente ligado entre los soportes C y D, y así sucesivamente hasta que todas las K uniones de soportes adyacentes estén ocupadas por un componente. En una reacción, se introducen componentes seleccionados de diferentes capas en los soportes junto con pares seleccionados de grapas que los dirigen para ensamblarse en los soportes apropiados. Por ejemplo, el par de grapas a^*L^* (5' a 3') y A^*b^* (5' a 3') dirigen el componente de la capa 1 con una región 'a' en el extremo 5' y una región 'b' en el extremo 3' para ligar entre los soportes L y A. En general, con M capas y $K+1$ soportes, se pueden usar 2^*M^K grapas seleccionables para construir cualquier identificador USS de longitud K . Debido a que las grapas que conectan un componente a un soporte en el extremo 5' están separadas de las grapas que conectan el mismo componente a un soporte en el extremo 3', se pueden formar subproductos de ácido nucleico en la reacción con soportes de bordes iguales como identificador diana, pero con menos de K componentes (menos de $K+1$ soportes) o con más de K componentes (más de $K+1$ soportes). El identificador diana puede formarse exactamente con K componentes ($K+1$ soportes) y, por lo tanto, puede ser seleccionable mediante técnicas como la selección del tamaño del ADN si todos los componentes están diseñados para tener la misma longitud y todos los soportes están diseñados para tener la misma longitud. En ciertas realizaciones del esquema de cadena sin restricciones donde puede haber un componente por capa, ese componente puede comprender únicamente una única secuencia de ácido nucleico distinta que cumple las tres funciones de (1) un código de barras de identificación, (2) una región de hibridación para ligación mediada por grapas del extremo 5' a un soporte, y (3) una región de hibridación para ligación mediada por grapas del extremo 3' a un soporte.

Los soportes internos ilustrados en la FIG. 14B pueden diseñarse de manera que utilicen la misma secuencia de hibridación tanto para la ligación en 5' mediada por grapas del soporte a un componente como para la ligación en 3' mediada por grapas del soporte a otro (no necesariamente distinto) componente. Por lo tanto, los eventos de hibridación apilados de un soporte y dos grapas representados en la FIG. 14B representan los eventos estadísticos de hibridación sucesivos que se producen entre el soporte y cada una de las grapas, permitiendo así tanto la ligación del componente 5' como la ligación del componente 3'. En otras realizaciones del esquema de cadena sin restricciones, el soporte puede diseñarse con dos regiones de hibridación concatenadas: una región de hibridación 3' distinta para la ligación 3' mediada por grapas y una región de hibridación 5' distinta para la ligación 5' mediada por grapas.

Las FIG. 15A y 15B ilustran esquemáticamente un método de ejemplo, denominado "esquema de eliminación de componentes", para construir identificadores mediante la delección de secuencias de ácidos nucleicos (o componentes) de un identificador parental. La FIG. 15A muestra un ejemplo de los espacios combinatorios de posibles identificadores que pueden construirse usando el esquema de eliminación de componentes. En este ejemplo, un identificador parental puede comprender múltiples componentes. Un identificador parental puede comprender más de o igual a aproximadamente 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 o más componentes.

Un identificador individual puede construirse al eliminar selectivamente cualquier número de componentes de N componentes posibles, conduciendo a un espacio combinatorio "completo" de tamaño 2^N , o al eliminar un número fijo de K componentes de N componentes posibles, conduciendo de esta manera a un espacio combinatorio " N choose K " de tamaño N choose K . En un ejemplo con un identificador parental con 3 componentes, el espacio combinatorio completo puede ser 8 y el espacio combinatorio 3choose2 puede ser 3.

La FIG. 15B muestra una implementación de ejemplo del esquema de eliminación de componentes usando escisión y reparación dirigidas a doble cadena (DSTCR). La secuencia parental puede ser un sustrato de ADN monocatenario que comprende componentes flanqueados por sitios diana específicos de nucleasa (que pueden tener 4 o menos bases de longitud), y donde el parental puede incubarse con una o más nucleasas específicas de doble cadena correspondientes a los sitios diana. Se puede seleccionar un componente individual para su eliminación con un ADN monocatenario complementario (o molde de escisión) que se une al ADN componente (y a los sitios de nucleasa flanqueantes) en el parental, formando así una secuencia bicatenaria estable en el parental que puede escindirse en ambos extremos por las nucleasas. Otro ADN monocatenario (o molde de reparación) se hibrida con los extremos disjuntos resultantes del parental (entre los cuales había estado la secuencia componente) y los reúne entre sí para la ligación, ya sea directamente o mediante un puente mediante una secuencia de reemplazo, de modo que las secuencias ligadas en el parental ya no contienen sitios activos diana de nucleasas. Nos referimos a este método como "Escisión Dirigida a Doble Cadena" (DSTC). La selección de tamaño se puede utilizar para seleccionar identificadores con un determinado número de componentes eliminados.

Alternativamente o, además, el identificador parental puede ser un sustrato de ácido nucleico bicatenario o monocatenario que comprende componentes separados por secuencias espaciadoras de modo que no haya dos componentes flanqueados por la misma secuencia. El identificador parental puede incubarse con la nucleasa Cas9. Se puede seleccionar un componente individual para su eliminación con ácidos ribonucleicos guía (los moldes de escisión) que se unen a los bordes del componente y permiten la escisión mediada por Cas9 en sus sitios flanqueantes. Un ácido nucleico monocatenario (el molde de reparación) puede hibridarse con los extremos disjuntos resultantes del identificador parental (p. ej., entre los extremos donde había estado la secuencia componente), uniéndolos así para la ligación. La ligación se puede realizar directamente o uniendo los extremos con una secuencia de reemplazo, de modo que las secuencias ligadas en el parental ya no contengan secuencias espaciadoras a las que Cas9 pueda dirigirse. Nos referimos a este método como "escisión y reparación dirigida específica de secuencia" o "SSTCR".

utilizar recombinasas para escindir, desplazar, invertir y transponer segmentos de ADN para crear diferentes moléculas de ácidos nucleicos. En general, con N recombinasas puede haber 2^N posibles identificadores contruidos a partir de un parental. En algunas realizaciones, se pueden organizar múltiples pares ortogonales de sitios de reconocimiento de diferentes recombinasas en un identificador parental de forma superpuesta de manera que la aplicación de una recombinasa afecte el tipo de evento de recombinación que ocurre cuando se aplica una recombinasa aguas abajo (véase Roquet et al., Synthetic recombinase-based state machines in living cells, Science 353 (6297): aad8559 (2016)). Un sistema de este tipo puede ser capaz de construir un identificador diferente para cada orden de N recombinasas, $N!$. Las recombinasas pueden ser de la familia de la tirosina, como Flp y Cre, o de la gran familia de las serina recombinasas, como PhiC31, Bxb1, TP901 o A118. El uso de recombinasas de la gran familia de serina recombinasas puede ser ventajoso porque facilitan la recombinación irreversible y, por lo tanto, pueden producir identificadores de manera más eficiente que otras recombinasas.

En algunos casos, una única secuencia de ácido nucleico se puede programar para convertirse en muchas secuencias de ácido nucleico distintas mediante la aplicación de numerosas recombinasas en un orden distinto. Se pueden generar aproximadamente e^M secuencias de ácido nucleico distintas aplicando M recombinasas en diferentes subconjuntos y órdenes de las mismas, cuando el número de recombinasas, M , puede ser menor de o igual a 7 para la gran familia de serina recombinasas. Cuando el número de recombinasas, M , puede ser mayor de 7, el número de secuencias que se pueden producir se aproxima a $3,9^M$, véase, p. ej., Roquet et al., Synthetic recombinase-based state machines in living cells, Science 353 (6297): aad8559 (2016). Los métodos adicionales para producir diferentes secuencias de ADN a partir de una secuencia común pueden incluir enzimas de edición de ácidos nucleicos dirigidas como CRISPR-Cas, TALENS y Nucleasas de Dedos de Cinc. Las secuencias producidas por recombinasas, enzimas de edición dirigida o similares se pueden usar junto con cualquiera de los métodos anteriores, por ejemplo, los métodos divulgados en cualquiera de las figuras y la divulgación en la presente solicitud.

Si la corriente de bits de información a codificar es mayor que la que puede codificarse mediante cualquier molécula de ácido nucleico individual, entonces la información se puede dividir e indexar con códigos de barras de secuencias de ácidos nucleicos. Además, cualquier subconjunto de moléculas de ácido nucleico de tamaño k del conjunto de N moléculas de ácido nucleico se puede elegir para producir $\log_2(N$ choose $k)$ bits de información. Se pueden ensamblar códigos de barras en las moléculas de ácido nucleico dentro de los subconjuntos de tamaño k para codificar corrientes de bits aún más largas. Por ejemplo, se pueden utilizar M códigos de barras para producir $M \cdot \log_2(N$ choose $k)$ bits de información. Dado un número, N , de moléculas de

ácido nucleico disponibles en un conjunto y un número, M , de códigos de barras disponibles, se pueden elegir subconjuntos de tamaño $k = k_0$ para minimizar el número total de moléculas en un combinado para codificar una pieza de información. Un método para codificar información digital puede comprender etapas para dividir la corriente de bits y codificar cada orden de N recombinasas, $N!$. Las recombinasas pueden ser de la familia de la tirosina, como Flp y Cre, o de la gran familia de las serina recombinasas, como PhiC31, Bxb1, TP901 o A118. El uso de recombinasas de la gran familia de serina recombinasas puede ser ventajoso porque facilitan la recombinación irreversible y, por lo tanto, pueden producir identificadores de manera más eficiente que otras recombinasas.

En algunos casos, una única secuencia de ácido nucleico se puede programar para convertirse en muchas secuencias de ácido nucleico distintas mediante la aplicación de numerosas recombinasas en un orden distinto. Se pueden generar aproximadamente $\sim e^M$ secuencias de ácido nucleico distintas aplicando M recombinasas en diferentes subconjuntos y órdenes de las mismas, cuando el número de recombinasas, M , puede ser menor de o igual a 7 para la gran familia de serina recombinasas. Cuando el número de recombinasas, M , puede ser mayor de 7, el número de secuencias que se pueden producir se aproxima a $3,9^M$, véase, p. ej., Roquet et al., Synthetic recombinase-based state machines in living cells, Science 353 (6297): aad8559 (2016), que se incorpora en su totalidad en la presente memoria por referencia. Los métodos adicionales para producir diferentes secuencias de ADN a partir de una secuencia común pueden incluir enzimas de edición de ácidos nucleicos dirigidas como CRISPR-Cas, TALENS y Nucleasas de Dedos de Cinc. Las secuencias producidas por recombinasas, enzimas de edición dirigida o similares se pueden usar junto con cualquiera de los métodos anteriores, por ejemplo, los métodos divulgados en cualquiera de las figuras y la divulgación en la presente solicitud.

Si la corriente de bits de información a codificar es mayor que el que puede codificarse mediante cualquier molécula de ácido nucleico individual, entonces la información se puede dividir e indexar con códigos de barras de secuencias de ácidos nucleicos. Además, cualquier subconjunto de moléculas de ácido nucleico de tamaño k del conjunto de N moléculas de ácido nucleico se puede elegir para producir $\log_2(N \text{ choose } k)$ bits de información. Se pueden ensamblar códigos de barras en las moléculas de ácido nucleico dentro de los subconjuntos de tamaño k para codificar corrientes de bits aún más largas. Por ejemplo, se pueden utilizar M códigos de barras para producir $M \cdot \log_2(N \text{ choose } k)$ bits de información. Dado un número, N , de moléculas de ácido nucleico disponibles en un conjunto y un número, M , de códigos de barras disponibles, se pueden elegir subconjuntos de tamaño $k = k_0$ para minimizar el número total de moléculas en un combinado para codificar una pieza de información. Un método para codificar información digital puede comprender etapas para dividir la corriente de bits y codificar los elementos individuales. Por ejemplo, una corriente de bits que comprende 6 bits se puede dividir en 3 componentes, comprendiendo cada componente dos bits. Cada componente de dos bits puede tener un código de barras para formar un casete de información y agruparse o combinarse para formar un hipercombinado de casetes de información.

Los códigos de barras pueden facilitar la indexación de información cuando la cantidad de información digital a codificar excede la cantidad que puede caber en un solo combinado. La información que comprende cadenas más largas de bits y/o múltiples bytes se puede codificar superponiendo el enfoque divulgado en la FIG. 3, por ejemplo, incluyendo una etiqueta con secuencias de ácido nucleico únicas codificadas utilizando el índice de ácido nucleico. Los casetes de información o bibliotecas de identificadores pueden comprender bases nitrogenadas o secuencias de ácido nucleico que incluyen secuencias de ácido nucleico únicas que proporcionan información de localización y valor de bits además de un código de barras o etiqueta que indica el componente o componentes de la corriente de bits a la que corresponde una secuencia determinada. Los casetes de información pueden comprender una o más secuencias de ácido nucleico únicas, así como un código de barras o etiqueta. El código de barras o etiqueta en el casete de información puede proporcionar una referencia para el casete de información y cualesquiera secuencias incluidas en el casete de información. Por ejemplo, la etiqueta o código de barras en un casete de información puede indicar para qué porción de la corriente de bits o componente de bits de la corriente de bits codifica información la secuencia única (p. ej., la información del valor del bit y posición del bit).

Utilizando códigos de barras, se puede codificar más información en bits en un combinado que el tamaño del espacio combinatorio de posibles identificadores. Una secuencia de 10 bits, por ejemplo, se puede separar en dos conjuntos de bytes, comprendiendo cada byte 5 bits. Cada byte se puede mapear a un conjunto de 5 posibles identificadores distintos. Inicialmente, los identificadores generados para cada byte pueden ser los mismos, pero pueden mantenerse en combinados separados o, de lo contrario, es posible que alguien que lea la información no pueda saber a qué byte pertenece una secuencia de ácido nucleico en particular. Sin embargo, cada identificador puede tener un código de barras o etiquetarse con un marcador que corresponda al byte al que se aplica la información codificada (p. ej., el código de barras uno puede unirse a secuencias en el combinado de ácidos nucleicos para proporcionar los primeros cinco bits y el código de barras dos puede unirse a las secuencias en el combinado de ácidos nucleicos para proporcionar los segundos cinco bits), y luego los identificadores correspondientes a los dos bytes se pueden combinar en un combinado (p. ej., "hipercombinado" o una o más bibliotecas de identificadores). Cada biblioteca de identificadores de una o más bibliotecas de identificadores combinadas puede comprender un código de barras distinto que identifica un

identificador determinado como perteneciente a una biblioteca de identificadores determinada. Los métodos para añadir un código de barras a cada identificador en una biblioteca de identificadores pueden comprender el uso de PCR, Gibson, ligación o cualquier otro enfoque que permita que un código de barras determinado (p. ej., el código de barras 1) se una a un combinado de muestras de ácido nucleico determinado (p. ej., el código de barras 1 al combinado de muestras de ácido nucleico 1 y el código de barras 2 al combinado de muestras de ácido nucleico 2). La muestra del hipercombinado se puede leer con métodos de secuenciación y la información de secuenciación se puede analizar usando el código de barras o la etiqueta. Un método que utiliza bibliotecas de identificadores y códigos de barras con un conjunto de M códigos de barras y N posibles identificadores (el espacio combinatorio) puede codificar una corriente de bits con una longitud equivalente al producto de M y N .

En algunas realizaciones, las bibliotecas de identificadores se pueden almacenar en una matriz de pocillos. La matriz de pocillos se puede definir como si tuviera n columnas y q filas y cada pocillo puede comprender dos o más bibliotecas de identificadores en un hipercombinado. La información codificada en cada pocillo puede constituir un gran elemento contiguo de información de tamaño $n \times q$ mayor que la información contenida en cada uno de los pocillos. Se puede tomar una alícuota de uno o más pocillos en la matriz de pocillos y se puede leer la codificación usando secuenciación, hibridación o PCR.

Un combinado de muestras de ácido nucleico, hipercombinado, biblioteca de identificadores, grupo de bibliotecas de identificadores, o un pocillo, que contiene un combinado o hipercombinado de muestras de ácido nucleico puede comprender moléculas de ácidos nucleicos únicas (p. ej., identificadores) correspondientes a bits de información y una pluralidad de secuencias de ácidos nucleicos suplementarias. Las secuencias de ácidos nucleicos suplementarias pueden no corresponder a datos codificados (p. ej., no corresponden a un valor de bit). Las muestras de ácido nucleico suplementarias pueden enmascarar o encriptar la información almacenada en el combinado de muestras. Las secuencias de ácido nucleico suplementarias pueden derivarse de una fuente biológica o producirse sintéticamente. Las secuencias de ácido nucleico suplementarias derivadas de una fuente biológica pueden incluir secuencias de ácido nucleico fragmentadas aleatoriamente o secuencias fragmentadas racionalmente. Los ácidos nucleicos suplementarios derivados biológicamente pueden ocultar u oscurecer los ácidos nucleicos que contienen datos dentro del combinado de muestras proporcionando información genética natural junto con la información codificada sintéticamente, especialmente si la información codificada sintéticamente (p. ej., el espacio combinatorio de identificadores) se hace para asemejar a la información genética natural (p. ej., un genoma fragmentado). En un ejemplo, los identificadores se derivan de una fuente biológica y los ácidos nucleicos suplementarios se derivan de una fuente biológica. Un combinado de muestras puede contener múltiples conjuntos de identificadores y secuencias de ácidos nucleicos suplementarias. Cada conjunto de identificadores y secuencias de ácidos nucleicos suplementarias puede derivarse de diferentes organismos. En un ejemplo, los identificadores se derivan de uno o más organismos y las secuencias de ácidos nucleicos suplementarias se derivan de un único organismo diferente. Las secuencias de ácidos nucleicos suplementarias también pueden derivarse de uno o más organismos y los identificadores pueden derivarse de un único organismo que es diferente del organismo del que se derivan los ácidos nucleicos suplementarios. Tanto los identificadores como las secuencias de ácidos nucleicos suplementarias pueden derivarse de múltiples organismos diferentes. Puede usarse una clave para distinguir los identificadores de las secuencias de ácidos nucleicos suplementarias.

Las secuencias de ácidos nucleicos suplementarias pueden almacenar metadatos sobre la información escrita. Los metadatos pueden comprender información adicional para determinar y/o autorizar la fuente de la información original y/o el destinatario previsto de la información original. Los metadatos pueden comprender información adicional sobre el formato de la información original, los instrumentos y métodos utilizados para codificar y escribir la información original, y la fecha y hora de escritura de la información original en los identificadores. Los metadatos pueden comprender información adicional sobre el formato de la información original, los instrumentos y métodos utilizados para codificar y escribir la información original, y la fecha y hora de escritura de la información original en secuencias de ácidos nucleicos. Los metadatos pueden comprender información adicional sobre modificaciones realizadas a la información original después de escribir la información en secuencias de ácidos nucleicos. Los metadatos pueden comprender anotaciones a la información original o una o más referencias a información externa. Alternativamente o, además, los metadatos pueden almacenarse en uno o más códigos de barras o etiquetas unidas a los identificadores.

Los identificadores en un combinado de identificadores pueden tener longitudes iguales, similares o diferentes entre sí. Las secuencias de ácidos nucleicos suplementarias pueden tener una longitud que es menor que, sustancialmente igual a o mayor que la longitud de los identificadores. Las secuencias de ácido nucleico suplementarias pueden tener una longitud promedio que está dentro de una base, dentro de dos bases, dentro de tres bases, dentro de cuatro bases, dentro de cinco bases, dentro de seis bases, dentro de siete bases, dentro de ocho bases, dentro de nueve bases, dentro de diez bases, o dentro de más bases de la longitud promedio de los identificadores. En un ejemplo, las secuencias de ácidos nucleicos suplementarias tienen la misma o sustancialmente la misma longitud que los identificadores. La concentración de secuencias de ácidos nucleicos suplementarias puede ser menor que, sustancialmente igual a o mayor que la concentración de los identificadores en la biblioteca de identificadores. La concentración de los ácidos nucleicos suplementarios

puede ser menor de o igual a aproximadamente el 1 %, 10 %, 20 %, 40 %, 60 %, 80 %, 100 %, 125 %, 150 %, 175 %, 200 %, 1.000 %, 1×10^4 %, 1×10^5 %, 1×10^6 %, 1×10^7 %, 1×10^8 % o menos que la concentración de los identificadores. La concentración de los ácidos nucleicos suplementarios puede ser mayor de o igual a aproximadamente el 1 %, 10 %, 20 %, 40 %, 60 %, 80 %, 100 %, 125 %, 150 %, 175 %, 200 %, 1.000 %, 1×10^4 %, 1×10^5 %, 1×10^6 %, 1×10^7 %, 1×10^8 % o más que la concentración de los identificadores. Las concentraciones más grandes pueden ser beneficiosas para ofuscar u ocultar datos. En un ejemplo, la concentración de las secuencias de ácidos nucleicos suplementarias es sustancialmente mayor (p. ej., 1×10^8 % mayor) que la concentración de identificadores en un combinado de identificadores.

10 Métodos para copiar y acceder a datos almacenados en secuencias de ácidos nucleicos

En otro aspecto, la presente divulgación proporciona métodos para copiar información codificada en secuencia(s) de ácidos nucleicos. Un método para copiar información codificada en secuencia(s) de ácidos nucleicos puede comprender (a) proporcionar una biblioteca de identificadores y (b) construir una o más copias de la biblioteca de identificadores. Una biblioteca de identificadores puede comprender un subconjunto de una pluralidad de identificadores de un espacio combinatorio más grande. Cada identificador individual de la pluralidad de identificadores puede corresponder a un símbolo individual en una cadena de símbolos. Un identificador puede comprender uno o más componentes. Un componente puede comprender una secuencia de ácido nucleico.

En otro aspecto, la presente divulgación proporciona métodos para acceder a la información codificada en secuencia(s) de ácidos nucleicos. Un método para acceder a información codificada en secuencias de ácidos nucleicos puede comprender (a) proporcionar una biblioteca de identificadores y (b) extraer una porción o un subconjunto de los identificadores presentes en la biblioteca de identificadores de la biblioteca de identificadores. Una biblioteca de identificadores puede comprender un subconjunto de una pluralidad de identificadores de un espacio combinatorio más grande. Cada identificador individual de la pluralidad de identificadores puede corresponder a un símbolo individual en una cadena de símbolos. Un identificador puede comprender uno o más componentes. Un componente puede comprender una secuencia de ácido nucleico.

La información se puede escribir en una o más bibliotecas de identificadores como se describe en otra parte de la presente memoria. Los identificadores pueden construirse utilizando cualquier método descrito en otra parte de la presente memoria. Los datos almacenados se pueden copiar generando copias de los identificadores individuales en una biblioteca de identificadores o en una o más bibliotecas de identificadores. Se puede copiar una parte de los identificadores o una biblioteca completa. La copia se puede realizar amplificando los identificadores en una biblioteca de identificadores. Cuando se combinan una o más bibliotecas de identificadores, se puede copiar una única biblioteca de identificadores o múltiples bibliotecas de identificadores. Si una biblioteca de identificadores comprende secuencias de ácidos nucleicos suplementarias, las secuencias de ácidos nucleicos suplementarias pueden copiarse o no.

Los identificadores en una biblioteca de identificadores pueden construirse para comprender uno o más sitios de unión de cebadores comunes. El uno o más sitios de unión pueden estar localizados en los bordes de cada identificador o intercalados a lo largo de cada identificador. El sitio de unión del cebador puede permitir que un par de cebadores específicos de la biblioteca de identificadores o un par de cebadores universales se unan a y amplifiquen los identificadores. Todos los identificadores dentro de una biblioteca de identificadores o todos los identificadores en una o más bibliotecas de identificadores pueden replicarse múltiples veces mediante múltiples ciclos de PCR. Puede usarse PCR convencional para copiar los identificadores y los identificadores pueden replicarse exponencialmente con cada ciclo de PCR. El número de copias de un identificador puede aumentar exponencialmente con cada ciclo de PCR. Puede usarse PCR lineal para copiar los identificadores y los identificadores pueden replicarse linealmente con cada ciclo de PCR. El número de copias del identificador puede aumentar linealmente con cada ciclo de PCR. Los identificadores pueden ligarse en un vector circular antes de la amplificación por PCR. El vector circular puede comprender un código de barras en cada extremo del sitio de inserción del identificador. Los cebadores de PCR para amplificar identificadores pueden diseñarse para cebarse con el vector de manera que los bordes con código de barras se incluyan con el identificador en el producto de amplificación. Durante la amplificación, la recombinación entre identificadores puede dar lugar a identificadores copiados que comprenden códigos de barras no correlacionados en cada borde. Los códigos de barras no correlacionados pueden ser detectables al leer los identificadores. Los identificadores que contienen códigos de barras no correlacionados pueden considerarse falsos positivos y pueden ignorarse durante el proceso de descodificación de la información.

La información se puede codificar asignando cada bit de información a una molécula de ácido nucleico única. Por ejemplo, tres conjuntos de muestras (X, Y y Z), conteniendo cada uno dos secuencias de ácido nucleico, pueden ensamblarse en ocho moléculas de ácido nucleico únicas y codificar ocho bits de datos:

N1 = X1Y1Z1

N2 = X1Y1Z2

N3 = X1Y2Z1

N4 = X1Y2Z2

5

N5 = X2Y1Z1

N6 = X2Y1Z2

10

N7 = X2Y2Z1

N8 = X2Y2Z2

15 Luego, cada bit en una cadena puede asignarse a la molécula de ácido nucleico correspondiente (p. ej., N1 puede especificar el primer bit, N2 puede especificar el segundo bit, N3 puede especificar el tercer bit, y así sucesivamente). La cadena de bits completa puede asignarse a una combinación de moléculas de ácido nucleico donde las moléculas de ácido nucleico correspondientes a valores de bits de '1' están incluidas en la combinación o combinado. Por ejemplo, en codificaciones UTF-8, la letra 'K' puede representarse mediante el código de cadena de 8 bits 01001011 que puede codificarse mediante la presencia de cuatro moléculas de ácido nucleico (p. ej., X1Y1Z2, X2Y1Z1, X2Y2Z1 y X2Y2Z2 en el ejemplo anterior).

25 Se puede acceder a la información mediante ensayos de secuenciación o hibridación. Por ejemplo, se pueden diseñar cebadores o sondas para unirse a regiones comunes o a la región con código de barras de la secuencia de ácido nucleico. Esto puede permitir la amplificación de cualquier región de la molécula de ácido nucleico. A continuación, el producto de amplificación puede leerse secuenciando el producto de amplificación o mediante un ensayo de hibridación. En el ejemplo anterior que codifica la letra 'K', si la primera mitad de los datos es de interés, se puede usar un cebador específico para la región del código de barras de la secuencia de ácido nucleico X1 y un cebador que se une a la región común del conjunto Z para amplificar las moléculas de ácido nucleico. Esto puede devolver la secuencia Y1Z2, que puede codificar 0100. También se puede acceder a la subcadena de esos datos amplificando aún más las moléculas de ácido nucleico con un cebador que se une a la región del código de barras de la secuencia de ácido nucleico Y1 y un cebador que se une a la secuencia común del conjunto Z. Esto puede devolver la secuencia de ácido nucleico Z2, que codifica la subcadena 01. Alternativamente, se puede acceder a los datos comprobando la presencia o ausencia de una secuencia de ácido nucleico particular sin secuenciar. Por ejemplo, la amplificación con un cebador específico para el código de barras Y2 puede generar productos de amplificación para el código de barras Y2, pero no para el código de barras Y1. La presencia del producto de amplificación Y2 puede indicar un valor de bit de '1'. Alternativamente, la ausencia de productos de amplificación Y2 puede indicar un valor de bit de '0'.

40 Se pueden usar métodos basados en PCR para acceder y copiar datos de identificadores o combinados de muestras de ácido nucleico. Utilizando sitios de unión de cebadores comunes que flanquean los identificadores en los combinados o hipercombinados, los ácidos nucleicos que contienen información se pueden copiar fácilmente. Alternativamente, también se pueden usar otros enfoques de amplificación de ácidos nucleicos, tales como la amplificación isotérmica, para copiar fácilmente datos de combinados o hipercombinados de muestras (p. ej., bibliotecas de identificadores). En los casos en los que la muestra comprende hipercombinados, se puede acceder y recuperar un subconjunto particular de información (p. ej., todos los ácidos nucleicos relacionados con un código de barras particular) usando un cebador que se une al código de barras específico en un borde del identificador en la orientación directa, junto con otro cebador que se une a una secuencia común en el borde opuesto del identificador en una orientación inversa. Se pueden utilizar varios métodos de lectura para extraer información del ácido nucleico codificado; por ejemplo, se pueden utilizar además micromatrices (o cualquier tipo de hibridación fluorescente), PCR digital, PCR cuantitativa (qPCR) y varias plataformas de secuenciación para leer las secuencias codificadas y, por extensión, los datos codificados digitalmente.

55 El acceso a la información almacenada en moléculas de ácido nucleico (p. ej., identificadores) se puede realizar eliminando selectivamente la porción de identificadores no diana de una biblioteca de identificadores o un combinado de identificadores o, por ejemplo, eliminando selectivamente todos los identificadores de una biblioteca de identificadores de un combinado de múltiples bibliotecas de identificadores. El acceso a los datos también se puede realizar capturando selectivamente identificadores diana de una biblioteca de identificadores o un combinado de identificadores. Los identificadores diana pueden corresponder a datos de interés dentro del elemento de información más amplio. Un combinado de identificadores puede comprender moléculas de ácido nucleico suplementarias. Las moléculas de ácido nucleico suplementarias pueden contener metadatos sobre la información codificada o pueden usarse para encriptar o enmascarar los identificadores correspondientes a la información. Las moléculas de ácido nucleico suplementarias pueden extraerse o no al acceder a los identificadores diana. Las FIG. 17A - 17C ilustran esquemáticamente una visión general de métodos de ejemplo para acceder a porciones de información almacenada en secuencias de ácido nucleico accediendo a un número de identificadores particulares desde un número mayor de identificadores. La FIG.

17A muestra métodos de ejemplo para usar la reacción en cadena de la polimerasa, sondas etiquetadas por afinidad y sondas dirigidas a la degradación para acceder a identificadores que contienen un componente específico. Para el acceso basado en PCR, un combinado de identificadores (p. ej., una biblioteca de identificadores) puede comprender identificadores con una secuencia común en cada extremo, una secuencia variable en cada extremo o una de una secuencia común o una secuencia variable en cada extremo. Las secuencias comunes o secuencias variables pueden ser sitios de unión de cebadores. Uno o más cebadores pueden unirse a las regiones comunes o variables en los bordes del identificador. Los identificadores con cebadores unidos pueden amplificarse mediante PCR. Los identificadores amplificados pueden superar significativamente a los identificadores no amplificados. Durante la lectura, se pueden identificar los identificadores amplificados. Un identificador de una biblioteca de identificadores puede comprender secuencias en uno o ambos de sus extremos que son distintivas de esa biblioteca, permitiendo así acceder selectivamente a una única biblioteca desde un combinado o grupo de más de una biblioteca de identificadores.

Para el acceso basado en etiquetas de afinidad, los componentes que constituyen los identificadores en un combinado pueden compartir complementariedad con una o más sondas. La una o más sondas pueden unirse o hibridarse con los identificadores a los que se accederá. La sonda puede comprender una etiqueta de afinidad. Las etiquetas de afinidad pueden unirse a una perla, generando un complejo que comprende una perla, al menos una sonda y al menos un identificador. Las perlas pueden ser magnéticas y, junto con un imán, las perlas pueden recoger y aislar los identificadores a los que se puede acceder. Los identificadores pueden eliminarse de las perlas en condiciones desnaturalizantes antes de la lectura. Alternativamente, o además, las perlas pueden recolectar los identificadores no diana y aislarlos del resto del combinado para que puedan lavarse en un recipiente separado y leerse. La etiqueta de afinidad puede unirse a una columna. Los identificadores a los que se accederá pueden unirse a la columna para su captura. Posteriormente, los identificadores unidos a la columna pueden eluirse o desnaturalizarse de la columna antes de la lectura. Alternativamente, los identificadores no diana pueden captarse selectivamente en la columna, mientras que los identificadores diana pueden fluir a través de la columna. El acceso a los identificadores diana puede comprender aplicar una o más sondas a un combinado de identificadores simultáneamente o aplicar una o más sondas a un combinado de identificadores secuencialmente.

Para el acceso basado en degradación, los componentes que constituyen los identificadores en un combinado pueden compartir complementariedad con una o más sondas dirigidas a la degradación. Las sondas pueden unirse o hibridarse con distintos componentes de los identificadores. La sonda puede ser una diana para una enzima de degradación, tal como una endonucleasa. En un ejemplo, se pueden combinar una o más bibliotecas de identificadores. Un conjunto de sondas puede hibridarse con una de las bibliotecas de identificadores. El conjunto de sondas puede comprender ARN y el ARN puede guiar una enzima Cas9. Se puede introducir una enzima Cas9 en una o más bibliotecas de identificadores. Los identificadores hibridados con las sondas pueden ser degradados por la enzima Cas9. Los identificadores a los que se accede no pueden ser degradados por la enzima de degradación. En otro ejemplo, los identificadores pueden ser monocatenarios y la biblioteca de identificadores puede combinarse con endonucleasas específicas de una cadena, tales como la nucleasa S1, que degrada selectivamente los identificadores a los que no se va a acceder. Los identificadores a los que se accede pueden hibridarse con un conjunto complementario de identificadores para protegerlos de la degradación por la o las endonucleasas específicas de una cadena. Los identificadores a los que se accede pueden separarse de los productos de degradación mediante selección de tamaño, tal como cromatografía de selección de tamaño (p. ej., electroforesis en gel de agarosa). Alternativamente o, además, los identificadores que no se degradan se pueden amplificar selectivamente (p. ej., usando PCR) de manera que los productos de degradación no se amplifiquen. Los identificadores no degradados se pueden amplificar usando cebadores que se hibridan con cada extremo de los identificadores no degradados y, por lo tanto, no con cada extremo de los identificadores degradados o escindidos.

La FIG. 17B muestra métodos de ejemplo para usar la reacción en cadena de la polimerasa para realizar operaciones 'O' o 'Y' para acceder a identificadores que contienen múltiples componentes. En un ejemplo, si dos cebadores directos se unen a conjuntos distintos de identificadores en el extremo izquierdo, entonces se puede lograr una amplificación 'O' de la unión de esos conjuntos de identificadores usando los dos cebadores directos juntos en una reacción de PCR múltiple con un cebador inverso que se une a todos los identificadores en el extremo derecho. En otro ejemplo, si un cebador directo se une a un conjunto de identificadores en el extremo izquierdo y un cebador inverso se une a un conjunto de identificadores en el extremo derecho, entonces se puede lograr una amplificación 'Y' de la intersección de esos dos conjuntos de identificadores usando el cebador directo y el cebador inverso juntos como un par de cebadores en una reacción de PCR.

La FIG. 17C muestra métodos de ejemplo para usar etiquetas de afinidad para realizar operaciones 'O' o 'Y' para acceder a identificadores que contienen múltiples componentes. En un ejemplo, si la sonda de afinidad 'P1' captura todos los identificadores con el componente 'C1' y otra sonda de afinidad 'P2' captura todos los identificadores con el componente 'C2', entonces el conjunto de todos los identificadores con C1 o C2 se puede capturar usando P1 y P2 simultáneamente (correspondiente a una operación 'O'). En otro ejemplo con los mismos componentes y sondas, el conjunto de todos los identificadores con C1 y C2 se puede capturar utilizando P1 y P2 secuencialmente (correspondiente a una operación 'Y').

Métodos para leer información almacenada en secuencias de ácido nucleico

5 En otro aspecto, la presente divulgación proporciona métodos para leer información codificada en secuencias de ácido nucleico. Un método para leer información codificada en secuencias de ácido nucleico puede comprender (a) proporcionar una biblioteca de identificadores, (b) identificar los identificadores presentes en la biblioteca de identificadores, (c) generar una cadena de símbolos a partir de los identificadores presentes en la biblioteca de identificadores, y (d) compilar información de la cadena de símbolos. Una biblioteca de identificadores puede comprender un subconjunto de una pluralidad de identificadores de un espacio combinatorio. Cada identificador individual del subconjunto de identificadores puede corresponder a un símbolo individual en una cadena de símbolos. Un identificador puede comprender uno o más componentes. Un componente puede comprender una secuencia de ácido nucleico.

10 en otra parte de la presente memoria. Los identificadores pueden construirse utilizando cualquier método descrito en otra parte de la presente memoria. Los datos almacenados se pueden copiar y acceder a ellos utilizando cualquier método descrito en otro lugar de la presente memoria.

15 El identificador puede comprender información relacionada con una localización del símbolo codificado, un valor del símbolo codificado, o tanto la localización como el valor del símbolo codificado. Un identificador puede incluir información relacionada con una localización del símbolo codificado y la presencia o ausencia del identificador en una biblioteca de identificadores puede indicar el valor del símbolo. La presencia de un identificador en una biblioteca de identificadores puede indicar un primer valor de símbolo (p. ej., primer valor de bit) en una cadena binaria y la ausencia de un identificador en una biblioteca de identificadores puede indicar un segundo valor de símbolo (p. ej., segundo valor de bit) en una cadena binaria. En un sistema binario, basar un valor de bit en la presencia o ausencia de un identificador en una biblioteca de identificadores puede reducir el número de identificadores ensamblados y, por lo tanto, reducir el tiempo de escritura. En un ejemplo, la presencia de un identificador puede indicar un valor de bit de '1' en la localización mapeada y la ausencia de un identificador puede indicar un valor de bit de '0' en la localización mapeada.

20 La generación de símbolos (p. ej., valores de bits) para una pieza de información puede incluir identificar la presencia o ausencia del identificador al que se puede asignar o codificar el símbolo (p. ej., bit). La determinación de la presencia o ausencia de un identificador puede incluir secuenciar los identificadores presentes o usar una matriz de hibridación para detectar la presencia de un identificador. En un ejemplo, la descodificación y lectura de las secuencias codificadas se puede realizar usando plataformas de secuenciación. Se describen ejemplos de plataformas de secuenciación en la Solicitud de Patente de EE. UU. No. de Ser. 14/465.685 presentada el 21 de agosto de 2014, Solicitud de Patente de EE. UU. No. de Ser. 13/886.234 presentada el 2 de mayo de 2013, y Solicitud de Patente de EE. UU. No. de Ser. 12/400.593 presentada el 9 de marzo de 2009.

30 En un ejemplo, la descodificación de datos codificados en ácido nucleico se puede lograr mediante secuenciación base por base de las cadenas de ácido nucleico, tal como secuenciación Illumina®, o utilizando una técnica de secuenciación que indique la presencia o ausencia de secuencias de ácido nucleico específicas, como el análisis de fragmentación mediante electroforesis capilar. La secuenciación puede emplear el uso de terminadores reversibles. La secuenciación puede emplear el uso de nucleótidos o análogos de nucleótidos naturales o no naturales (p. ej., modificados por ingeniería). Alternativamente o además, la descodificación de secuencias de ácido nucleico se puede realizar usando una variedad de técnicas analíticas, que incluyen, pero no se limitan a, cualquier método que genere señales ópticas, electroquímicas o químicas. Se puede usar una variedad de enfoques de secuenciación que incluyen, pero no se limitan a, reacción en cadena de la polimerasa (PCR), PCR digital, secuenciación de Sanger, secuenciación de alto rendimiento, secuenciación por síntesis, secuenciación de una sola molécula, secuenciación por ligación, RNA-Seq (Illumina), secuenciación de próxima generación, expresión génica digital (Helicos), micromatriz única clonal (Solexa), secuenciación de escopeta, secuenciación Maxim-Gilbert o secuenciación masiva paralela. secuenciación de una sola molécula, secuenciación por ligación, RNA-Seq (Illumina), secuenciación de próxima generación, expresión génica digital (Helicos), micromatriz única clonal (Solexa), secuenciación de escopeta, secuenciación Maxim-Gilbert o secuenciación masiva paralela.

40 Se pueden usar varios métodos de lectura para extraer información del ácido nucleico codificado. En un ejemplo, se pueden utilizar además micromatrices (o cualquier tipo de hibridación fluorescente), PCR digital, PCR cuantitativa (qPCR) y varias plataformas de secuenciación para leer las secuencias codificadas y, por extensión, los datos codificados digitalmente.

50 Una biblioteca de identificadores puede comprender además secuencias de ácido nucleico suplementarias que proporcionan metadatos sobre la información, encriptan o enmascaran la información, o que proporcionan metadatos y enmascaran la información. Los ácidos nucleicos suplementarios pueden identificarse simultáneamente con la identificación de los identificadores. Alternativamente, los ácidos nucleicos suplementarios pueden identificarse antes o después de identificar los identificadores. En un ejemplo, los

ácidos nucleicos suplementarios no se identifican durante la lectura de la información codificada. Las secuencias de ácido nucleico suplementarias pueden ser indistinguibles de los identificadores. Puede usarse un índice de identificador o una clave para diferenciar las moléculas de ácido nucleico suplementarias de los identificadores.

5

La eficiencia de la codificación y decodificación de los datos se puede aumentar recodificando cadenas de bits de entrada para permitir el uso de menos moléculas de ácido nucleico. Por ejemplo, si se recibe una cadena de entrada con una alta aparición de subcadenas '111', que pueden asignarse a tres moléculas de ácido nucleico (p. ej., identificadores) con un método de codificación, se puede recodificar en una subcadena '000' que puede asignarse a un conjunto nulo de moléculas de ácido nucleico. La subcadena de entrada alternativa de '000' también se puede recodificar en '111'. Este método de recodificación puede reducir la cantidad total de moléculas de ácido nucleico usadas para codificar los datos porque puede haber una reducción en el número de '1' en el conjunto de datos. En este ejemplo, el tamaño total del conjunto de datos se puede aumentar para dar cabida a un libro de códigos que especifica las nuevas instrucciones de mapeo. Un método alternativo para aumentar la eficiencia de codificación y decodificación puede ser recodificar la cadena de entrada para reducir la longitud variable. Por ejemplo, '111' puede recodificarse como '00', lo que puede reducir el tamaño del conjunto de datos y reducir el número de '1' en el conjunto de datos.

10

15

20

25

30

35

La velocidad y eficiencia de decodificar datos codificados en ácidos nucleicos se pueden controlar (p. ej., aumentar) diseñando específicamente identificadores para facilitar la detección. Por ejemplo, las secuencias de ácido nucleico (p. ej., identificadores) que están diseñadas para facilitar la detección pueden incluir secuencias de ácido nucleico que comprenden una mayoría de nucleótidos que son más fáciles de llamar y detectar en función de sus propiedades ópticas, electroquímicas, químicas o físicas. Las secuencias de ácido nucleico diseñadas por ingeniería pueden ser monocatenarias o bicatenarias. Las secuencias de ácido nucleico diseñadas por ingeniería pueden incluir nucleótidos sintéticos o no naturales que mejoran las propiedades detectables de la secuencia de ácido nucleico. Las secuencias de ácido nucleico diseñadas por ingeniería pueden comprender todos los nucleótidos naturales, todos los nucleótidos sintéticos o no naturales, o una combinación de nucleótidos naturales, sintéticos y no naturales. Los nucleótidos sintéticos pueden incluir análogos de nucleótidos tales como ácidos nucleicos peptídicos, ácidos nucleicos bloqueados, ácidos nucleicos de glicol y ácidos nucleicos de treosa. Los nucleótidos no naturales pueden incluir dNaM, un nucleósido artificial que contiene un grupo 3-metoxi-2-naftilo, y d5SICS, un nucleósido artificial que contiene un grupo 6-metilisoquinolina-1-tiona-2-ilo. Las secuencias de ácido nucleico diseñadas por ingeniería pueden diseñarse para una única propiedad mejorada, tal como propiedades ópticas mejoradas, o las secuencias de ácido nucleico diseñadas pueden diseñarse con múltiples propiedades mejoradas, tales como propiedades ópticas y electroquímicas mejoradas o propiedades ópticas y químicas mejoradas.

40

45

Las secuencias de ácido nucleico diseñadas por ingeniería pueden comprender nucleótidos reactivos naturales, sintéticos y no naturales que no mejoran las propiedades ópticas, electroquímicas, químicas o físicas de las secuencias de ácido nucleico. Los componentes reactivos de las secuencias de ácido nucleico pueden permitir la adición de un resto químico que confiera propiedades mejoradas a la secuencia de ácido nucleico. Cada secuencia de ácido nucleico puede incluir un único resto químico o puede incluir múltiples restos químicos. Los restos químicos de ejemplo pueden incluir, pero no se limitan a, restos fluorescentes, restos quimioluminiscentes, restos ácidos o básicos, restos hidrófobos o hidrófilos y restos que alteran el estado de oxidación o la reactividad de la secuencia de ácido nucleico.

50

55

60

65

Se puede diseñar una plataforma de secuenciación específicamente para decodificar y leer información codificada en secuencias de ácido nucleico. La plataforma de secuenciación puede estar especializada en secuenciar moléculas de ácido nucleico monocatenarias o bicatenarias. La plataforma de secuenciación puede decodificar datos codificados en ácido nucleico leyendo bases individuales (p. ej., secuenciación base por base) o detectando la presencia o ausencia de una secuencia de ácido nucleico completa (p. ej., componente) incorporada dentro de la molécula de ácido nucleico (p. ej., identificador). La plataforma de secuenciación puede incluir el uso de reactivos promiscuos, mayores longitudes de lectura y la detección de secuencias de ácido nucleico específicas mediante la adición de restos químicos detectables. El uso de reactivos más promiscuos durante la secuenciación puede aumentar la eficiencia de la lectura al permitir una llamada de base más rápida, lo que a su vez puede disminuir el tiempo de secuenciación. El uso de longitudes de lectura aumentadas puede permitir decodificar secuencias más largas de ácidos nucleicos codificados por lectura. La adición de etiquetas de restos químicos detectables puede permitir la detección de la presencia o ausencia de una secuencia de ácido nucleico mediante la presencia o ausencia de un resto químico. Por ejemplo, cada secuencia de ácido nucleico que codifica un bit de información puede etiquetarse con un resto químico que genera una señal óptica, electroquímica o química única. La presencia o ausencia de esa señal óptica, electroquímica o química única puede indicar un valor de bit de '0' o '1'. La secuencia de ácido nucleico puede comprender un único resto químico o múltiples restos químicos. El resto químico puede añadirse a la secuencia de ácido nucleico antes del uso de la secuencia de ácido nucleico para codificar datos. Alternativamente o además, el resto químico puede añadirse a la secuencia de ácido nucleico después de codificar los datos, pero antes de decodificar los datos. La etiqueta del resto químico se puede añadir directamente a la secuencia de ácido nucleico o la secuencia de ácido nucleico puede comprender un anclaje de nucleótido sintético o no

natural y la etiqueta del resto químico se puede añadir a ese anclaje.

Se pueden aplicar códigos únicos para minimizar o detectar errores de codificación y decodificación. Pueden producirse errores de codificación y decodificación debido a falsos negativos (p. ej., una molécula de ácido nucleico o un identificador no incluido en un muestreo aleatorio). Un ejemplo de un código de detección de errores puede ser una secuencia de suma de comprobación que cuenta el número de identificadores en un conjunto contiguo de posibles identificadores que se incluye en la biblioteca de identificadores. Mientras se lee la biblioteca de identificadores, la suma de comprobación puede indicar cuántos identificadores de ese conjunto contiguo de identificadores se espera recuperar, y se pueden seguir muestreando los identificadores para su lectura hasta que se alcance el número esperado. En algunas realizaciones, se puede incluir una secuencia de suma de comprobación para cada conjunto contiguo de R identificadores donde R puede ser igual en tamaño o mayor de 1, 2, 5, 10, 50, 100, 200, 500 o 1.000 o menor de 1.000, 500, 200, 100, 50, 10, 5 o 2. Cuanto menor sea el valor de R , mejor será la detección de errores. En algunas realizaciones, las sumas de comprobación pueden ser secuencias de ácido nucleico suplementarias. Por ejemplo, un conjunto que comprende siete secuencias de ácido nucleico (p. ej., componentes) se puede dividir en dos grupos, secuencias de ácido nucleico para construir identificadores con un esquema de producto (componentes X1-X3 en la capa X e Y1-Y3 en la capa Y), y secuencias de ácido nucleico para las sumas de comprobación suplementarias (X4-X7 e Y4-Y7). Las secuencias de suma de comprobación X4-X7 pueden indicar si cero, una, dos o tres secuencias de la capa X están ensambladas con cada miembro de la capa Y. Alternativamente, las secuencias de suma de comprobación Y4-Y7 pueden indicar si cero, una, dos o tres secuencias de la capa Y están ensambladas con cada miembro de la capa X. En este ejemplo, una biblioteca de identificadores original con identificadores {X1Y1, X1Y3, X2Y1, X2Y2, X2Y3} se puede suplementar para incluir sumas de comprobación para convertirse en el siguiente combinado: {X1Y1, X1Y3, X2Y1, X2Y2, X2Y3, X1Y6, X2Y7, X3Y4, X6Y1, X5Y2, X6Y3}. Las secuencias de suma de comprobación también se pueden utilizar para la corrección de errores. Por ejemplo, la ausencia de X1Y1 en el conjunto de datos anterior y la presencia de X1Y6 y X6Y1 pueden permitir inferir que la molécula de ácido nucleico X1Y1 falta en el conjunto de datos. Las secuencias de suma de comprobación pueden indicar si faltan identificadores en un muestreo de la biblioteca de identificadores o en una parte de la biblioteca de identificadores a la que se ha accedido. En el caso de que falte una secuencia de suma de comprobación, los métodos de acceso como la PCR o la hibridación de sondas etiquetadas por afinidad pueden amplificarla y/o aislarla. En algunas realizaciones, las sumas de comprobación pueden ser secuencias de ácido nucleico suplementarias. Las sumas de verificación pueden codificarse directamente en la información de manera que estén representadas por identificadores.

El ruido en la codificación y decodificación de los datos se puede reducir mediante la construcción de identificadores de forma palindrómica, por ejemplo, mediante el uso de pares palindrómicos de componentes en lugar de componentes individuales en el esquema del producto. Luego, los pares de componentes de diferentes capas pueden ensamblarse entre sí de manera palindrómica (p. ej., YXY en lugar de XY para los componentes X e Y). Este método palindrómico puede expandirse a un mayor número de capas (p. ej., ZYXYZ en lugar de XYZ) y puede permitir la detección de reacciones cruzadas erróneas entre identificadores.

La adición de secuencias de ácido nucleico suplementarias en exceso (p. ej., gran exceso) a los identificadores puede evitar que la secuenciación recupere los identificadores codificados. Antes de decodificar la información, los identificadores pueden enriquecerse a partir de las secuencias de ácido nucleico suplementarias. Por ejemplo, los identificadores pueden enriquecerse mediante una reacción de amplificación de ácidos nucleicos usando cebadores específicos para los extremos del identificador. Alternativamente, o además, la información puede decodificarse sin enriquecer el combinado de muestras mediante secuenciación (p. ej., secuenciación por síntesis) usando un cebador específico. En ambos métodos de decodificación, puede resultar difícil enriquecer o decodificar la información sin tener una clave de decodificación o saber algo sobre la composición de los identificadores. También se pueden emplear métodos de acceso alternativos, como el uso de sondas basadas en etiquetas de afinidad.

Sistemas para codificar datos de secuencia binaria

Un sistema para codificar información digital en ácidos nucleicos (p. ej., ADN) puede comprender sistemas, métodos y dispositivos para convertir archivos y datos (p. ej., datos sin procesar, archivos zip comprimidos, datos enteros y otras formas de datos) en bytes y codificar los bytes en segmentos o secuencias de ácidos nucleicos, típicamente ADN, o combinaciones de los mismos.

En un aspecto, la presente divulgación proporciona sistemas para codificar datos de secuencia binaria usando ácidos nucleicos. Un sistema para codificar datos de secuencias binarias usando ácidos nucleicos puede comprender un dispositivo y uno o más procesadores informáticos. El dispositivo puede configurarse para construir una biblioteca de identificadores. Uno o más procesadores informáticos pueden programarse individual o colectivamente para (i) traducir la información en una cadena de símbolos, (ii) mapear la cadena de símbolos a la pluralidad de identificadores, y (iii) construir una biblioteca de identificadores que comprende al menos un subconjunto de una pluralidad de identificadores. Un identificador individual de la pluralidad de identificadores puede corresponder a un símbolo individual de una cadena de símbolos. Un identificador

individual de la pluralidad de identificadores puede comprender uno o más componentes. Un componente individual del uno o más componentes puede comprender una secuencia de ácido nucleico.

En otro aspecto, la presente divulgación proporciona sistemas para leer datos de secuencia binaria usando ácidos nucleicos. Un sistema para leer datos de secuencias binarias usando ácidos nucleicos puede comprender una base de datos y uno o más procesadores informáticos. La base de datos puede almacenar una biblioteca de identificadores que codifica la información. El uno o más procesadores informáticos pueden programarse individual o colectivamente para (i) identificar los identificadores en la biblioteca de identificadores, (ii) generar una pluralidad de símbolos a partir de identificadores identificados en (i), y (iii) compilar la información de la pluralidad de símbolos. La biblioteca de identificadores puede comprender un subconjunto de una pluralidad de identificadores. Cada identificador individual de la pluralidad de identificadores puede corresponder a un símbolo individual en una cadena de símbolos. Un identificador puede comprender uno o más componentes. Un componente puede comprender una secuencia de ácido nucleico.

Las realizaciones no limitantes de métodos para usar el sistema para codificar datos digitales pueden comprender etapas para recibir información digital en forma de corrientes de bytes. Análisis de las corrientes de bytes en bytes individuales, mapear la localización de un bit dentro del byte usando un índice de ácido nucleico (o rango de identificador) y codificar secuencias correspondientes a valores de bits de 1 o valores de bits de 0 en identificadores. Las etapas para recuperar datos digitales pueden comprender secuenciar una muestra de ácido nucleico o un combinado de ácidos nucleicos que comprende secuencias de ácido nucleico (p. ej., identificadores) que se mapean a uno o más bits, hacer referencia a un rango de identificador para confirmar si el identificador está presente en el combinado de ácidos nucleicos y descodificar la información de localización y valor de bits para cada secuencia en un byte que comprende una secuencia de información digital.

Los sistemas para codificar, escribir, copiar, acceder, leer y descodificar información codificada y escrita en moléculas de ácido nucleico pueden ser una única unidad integrada o pueden ser múltiples unidades configuradas para ejecutar una o más de las operaciones mencionadas anteriormente. Un sistema para codificar y escribir información en moléculas de ácido nucleico (p. ej., identificadores) puede incluir un dispositivo y uno o más procesadores informáticos. El uno o más procesadores informáticos pueden programarse para analizar la información en cadenas de símbolos (p. ej., cadenas de bits). El procesador informático puede generar un rango de identificador. El procesador informático puede clasificar los símbolos en dos o más categorías. Una categoría puede incluir símbolos que se representarán mediante la presencia del identificador correspondiente en la biblioteca de identificadores y la otra categoría puede incluir símbolos que se representarán mediante la ausencia de los identificadores correspondientes en la biblioteca de identificadores. El procesador informático puede dirigir al dispositivo para ensamblar los identificadores correspondientes a los símbolos que se van a representar en presencia de un identificador en la biblioteca de identificadores.

El dispositivo puede comprender una pluralidad de regiones, secciones o particiones. Los reactivos y componentes para ensamblar los identificadores pueden almacenarse en una o más regiones, secciones o particiones del dispositivo. Las capas pueden almacenarse en regiones separadas de la sección del dispositivo. Una capa puede comprender uno o más componentes únicos. El componente de una capa puede ser único respecto de los componentes de otra capa. Las regiones o secciones pueden comprender recipientes y las particiones pueden comprender pozos. Cada capa puede almacenarse en un recipiente o partición separada. Cada reactivo o secuencia de ácido nucleico puede almacenarse en un recipiente o partición separada. Alternativamente o, además, se pueden combinar reactivos para formar una mezcla maestra para la construcción de identificadores. El dispositivo puede transferir reactivos, componentes y moldes de una sección del dispositivo para combinarlos en otra sección. El dispositivo puede proporcionar las condiciones para completar la reacción de ensamblaje. Por ejemplo, el dispositivo puede proporcionar calentamiento, agitación y detección del progreso de la reacción. Los identificadores construidos pueden dirigirse para experimentar una o más reacciones posteriores para añadir códigos de barras, secuencias comunes, secuencias variables o etiquetas a uno o más extremos de los identificadores. Luego, los identificadores pueden dirigirse a una región o partición para generar una biblioteca de identificadores. Se pueden almacenar una o más bibliotecas de identificadores en cada región, sección o partición individual del dispositivo. El dispositivo puede transferir fluido (p. ej., reactivos, componentes, moldes) mediante presión, vacío o succión.

Las bibliotecas de identificadores pueden almacenarse en el dispositivo o pueden moverse a una base de datos separada. La base de datos puede comprender una o más bibliotecas de identificadores. La base de datos puede proporcionar condiciones para el almacenamiento a largo plazo de las bibliotecas de identificadores (p. ej., condiciones para reducir la degradación de los identificadores). Las bibliotecas de identificadores pueden almacenarse en forma de polvo, líquido o sólido. La base de datos puede proporcionar protección contra la luz ultravioleta, temperatura reducida (p. ej., refrigeración o congelación) y protección contra enzimas y sustancias químicas degradantes. Antes de transferirse a una base de datos, las bibliotecas de identificadores pueden liofilizarse o congelarse. Las bibliotecas de identificadores pueden incluir ácido etilendiaminotetraacético (EDTA) para inactivar nucleasas y/o un tampón para mantener la estabilidad de las moléculas de ácido

nucleico.

La base de datos puede estar acoplada, incluir o estar separada de un dispositivo que escribe la información en identificadores, copia la información, accede a la información o lee la información. Una parte de una biblioteca de identificadores puede eliminarse de la base de datos antes de copiarla, acceder a ella o leerla. El dispositivo que copia la información de la base de datos puede ser el mismo dispositivo o uno diferente del que escribe la información. El dispositivo que copia la información puede extraer una alícuota de una biblioteca de identificadores del dispositivo y combinar esa alícuota con los reactivos y constituyentes para amplificar una parte o toda la biblioteca de identificadores. El dispositivo puede controlar la temperatura, presión y agitación de la reacción de amplificación. El dispositivo puede comprender particiones y pueden ocurrir una o más reacciones de amplificación en la partición que comprende la biblioteca de identificadores. El dispositivo puede copiar más de un combinado de identificadores a la vez.

Los identificadores copiados pueden transferirse desde el dispositivo de copia a un dispositivo de acceso. El dispositivo de acceso puede ser el mismo dispositivo que el dispositivo de copia. El dispositivo de acceso puede comprender regiones, secciones o particiones separadas. El dispositivo de acceso puede tener una o más columnas, depósitos de perlas o regiones magnéticas para separar identificadores unidos a etiquetas de afinidad. Alternativamente, o además, el dispositivo de acceso puede tener una o más unidades de selección de tamaño. Una unidad de selección de tamaño puede incluir electroforesis en gel de agarosa o cualquier otro método para seleccionar el tamaño de moléculas de ácido nucleico. La copia y la extracción se pueden realizar en la misma región de un dispositivo o en diferentes regiones de un dispositivo.

Los datos a los que se accede se pueden leer en el mismo dispositivo o los datos a los que se accede se pueden transferir a otro dispositivo. El dispositivo de lectura puede comprender una unidad de detección para detectar e identificar los identificadores. La unidad de detección puede ser parte de un secuenciador, matriz de hibridación u otra unidad para identificar la presencia o ausencia de un identificador. Se puede diseñar una plataforma de secuenciación específicamente para descodificar y leer información codificada en secuencias de ácido nucleico. La plataforma de secuenciación puede estar especializada en secuenciar moléculas de ácido nucleico monocatenarias o bicatenarias. La plataforma de secuenciación puede descodificar datos codificados en ácido nucleico leyendo bases individuales (p. ej., secuenciación base por base) o detectando la presencia o ausencia de una secuencia de ácido nucleico completa (p. ej., componente) incorporada dentro de la molécula de ácido nucleico (p. ej., identificador). Alternativamente, la plataforma de secuenciación puede ser un sistema como Secuenciación Illumina® o análisis de fragmentación mediante electroforesis capilar. Alternativamente o además, la descodificación de secuencias de ácido nucleico se puede realizar usando una variedad de técnicas analíticas implementadas por el dispositivo, que incluyen, pero no se limitan a, cualesquiera de los métodos que generen señales ópticas, electroquímicas o químicas.

El almacenamiento de información en moléculas de ácido nucleico puede tener diversas aplicaciones que incluyen, pero no se limitan a, almacenamiento de información a largo plazo, almacenamiento de información sensible y almacenamiento de información médica. En un ejemplo, la información médica de una persona (p. ej., historial y registros médicos) puede almacenarse en moléculas de ácido nucleico y llevarse encima. La información puede almacenarse de manera externa al cuerpo (p. ej., en un dispositivo portátil) o interna al cuerpo (p. ej., en una cápsula subcutánea). Cuando se lleva a un paciente a un consultorio médico u hospital, se puede tomar una muestra del dispositivo o cápsula y la información se puede descodificar con el uso de un secuenciador de ácido nucleico. El almacenamiento personal de registros médicos en moléculas de ácido nucleico puede proporcionar una alternativa a los sistemas de almacenamiento basados en computadoras y en la nube. El almacenamiento personal de registros médicos en moléculas de ácido nucleico puede reducir la incidencia o la prevalencia de acceso ilegal a registros médicos. Las moléculas de ácido nucleico utilizadas para el almacenamiento de registros médicos en cápsulas pueden derivarse de secuencias genómicas humanas. El uso de secuencias genómicas humanas puede disminuir la inmunogenicidad de las secuencias de ácido nucleico en el caso de fallo y fuga de la cápsula.

Sistemas informáticos

La presente divulgación proporciona sistemas informáticos que están programados para implementar métodos de la divulgación. La FIG. 19 muestra un sistema informático **1901** que está programado o configurado de otro modo para codificar información digital en secuencias de ácido nucleico y/o leer (p. ej., descodificar) información derivada de secuencias de ácido nucleico. El sistema informático **1901** puede regular varios aspectos de los procedimientos de codificación y descodificación de la presente divulgación, como, por ejemplo, los valores de bits y la información de localización de bits para un bit o byte determinado de una corriente de bits o corriente de bytes codificada.

El sistema informático **1901** incluye una unidad central de procesamiento (CPU, también "procesador" y "procesador informático" en la presente memoria) **1905**, que puede ser un procesador de un solo núcleo o de múltiples núcleos, o una pluralidad de procesadores para procesamiento paralelo. El sistema informático **1901** también incluye memoria o localización de memoria **1910** (p. ej., memoria de acceso aleatorio, memoria solo

de lectura, memoria flash), unidad de almacenamiento electrónico **1915** (p. ej., disco duro), interfaz de comunicación **1920** (p. ej., adaptador de red) para comunicarse con uno o más de otros sistemas y dispositivos periféricos **1925**, tales como caché, otra memoria, almacenamiento de datos y/o adaptadores de pantalla electrónica. La memoria **1910**, la unidad de almacenamiento **1915**, la interfaz **1920** y los dispositivos periféricos **1925** están en comunicación con la CPU **1905** a través de un bus de comunicación (líneas continuas), tal como una placa base. La unidad de almacenamiento **1915** puede ser una unidad de almacenamiento de datos (o depósito de datos) para almacenar datos. El sistema informático **1901** puede acoplarse operativamente a una red informática ("red") **1930** con la ayuda de la interfaz de comunicación **1920**. La red **1930** puede ser Internet, una internet y/o extranet, o una intranet y/o extranet que está en comunicación con Internet. La red **1930** en algunos casos es una red de telecomunicaciones y/o datos. La red **1930** puede incluir uno o más servidores informáticos, que pueden permitir la computación distribuida, como la computación en la nube. La red **1930**, en algunos casos con la ayuda del sistema informático **1901**, puede implementar una red de igual a igual, que puede permitir que los dispositivos acoplados al sistema informático **1901** se comporten como un cliente o un servidor.

La CPU **1905** puede ejecutar una secuencia de instrucciones legibles por máquina, que pueden incorporarse en un programa o software. Las instrucciones pueden almacenarse en una localización de memoria, como la memoria **1910**. Las instrucciones pueden dirigirse a la CPU **1905**, que posteriormente puede programar o configurar de otro modo la CPU **1905** para implementar métodos de la presente divulgación. Los ejemplos de operaciones realizadas por la CPU **1905** pueden incluir buscar, descodificar, ejecutar y reescribir.

La CPU **1905** puede ser parte de un circuito, tal como un circuito integrado. Uno o más de otros componentes del sistema **1901** pueden incluirse en el circuito. En algunos casos, el circuito es un circuito integrado de aplicación específica (ASIC).

La unidad de almacenamiento **1915** puede almacenar archivos, tales como controladores, bibliotecas y programas guardados. La unidad de almacenamiento **1915** puede almacenar datos de usuario, p. ej., preferencias de usuario y programas de usuario. El sistema informático **1901** en algunos casos puede incluir una o más unidades de almacenamiento de datos adicionales que son externas al sistema informático **1901**, como las localizadas en un servidor remoto que está en comunicación con el sistema informático **1901** a través de una intranet o Internet.

El sistema informático **1901** puede comunicarse con uno o más sistemas informáticos remotos a través de la red **1930**. Por ejemplo, el sistema informático **1901** puede comunicarse con un sistema informático remoto de un usuario u otros dispositivos y/o maquinaria que puede ser utilizada por el usuario en el curso del análisis de datos codificados o descodificados en una secuencia de ácidos nucleicos. ácidos (p. ej., un secuenciador u otro sistema para determinar químicamente el orden de las bases nitrogenadas en una secuencia de ácido nucleico). Los ejemplos de sistemas informáticos remotos incluyen computadoras personales (p. ej., PC portátiles), pizarras o tabletas (p. ej., iPad de Apple®, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (p. ej., iPhone de Apple®, dispositivos con Android, BlackBerry®), o asistentes digitales personales. El usuario puede acceder al sistema informático **1901** a través de la red **1930**.

Los métodos como se describen en la presente memoria se pueden implementar mediante un código ejecutable de máquina (p. ej., procesador informático) almacenado en una localización de almacenamiento electrónico del sistema informático **1901**, tal como, por ejemplo, en la memoria **1910** o la unidad de almacenamiento electrónico **1915**. El código ejecutable por máquina o legible por máquina puede proporcionarse en forma de software. Durante el uso, el código puede ser ejecutado por el procesador **1905**. En algunos casos, el código puede recuperarse de la unidad de almacenamiento **1915** y almacenarse en la memoria **1910** para que el procesador **1905** pueda acceder fácilmente a él. En algunas situaciones, se puede excluir la unidad de almacenamiento electrónico **1915** y las instrucciones ejecutables por máquina se almacenan en la memoria **1910**.

El código se puede precompilar y configurar para su uso con una máquina que tenga un procesador adaptado para ejecutar el código, o se puede compilar durante el tiempo de ejecución. El código se puede suministrar en un lenguaje de programación que se puede seleccionar para permitir que el código se ejecute de forma precompilada o tal como está compilado.

Los aspectos de los sistemas y métodos proporcionados en la presente memoria, como el sistema informático **1901**, se pueden incorporar en la programación. Diversos aspectos de la tecnología se pueden considerar como "productos" o "artículos de fabricación" típicamente en la forma de código ejecutable por máquina (o procesador) y/o datos asociados que se portan o se incorporan en un tipo de medio legible por máquina. El código ejecutable por máquina se puede almacenar en una unidad de almacenamiento electrónico, tal como una memoria (p. ej., memoria solo de lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo "almacenamiento" pueden incluir cualquiera o toda la memoria tangible de los ordenadores, procesadores o similares, o módulos asociados de los mismos, tales como varias memorias semiconductoras, unidades de cinta, unidades de disco y similares, que pueden proporcionar almacenamiento no transitorio en

cualquier momento para la programación de software. En ocasiones, todo o parte del software puede comunicarse a través de Internet o de otras redes de telecomunicaciones. Dichas comunicaciones, por ejemplo, pueden permitir la carga del software desde una computadora o procesador a otro, por ejemplo, desde un servidor de administración o computadora central a la plataforma informática de un servidor de aplicaciones.

5 Por lo tanto, otro tipo de medio que puede soportar los elementos de software incluye ondas ópticas, eléctricas y electromagnéticas, tales como las utilizadas a través de interfaces físicas entre dispositivos locales, a través de redes terrestres cableadas y ópticas y a través de diversos enlaces aéreos. Los elementos físicos que transportan dichas ondas, tales como enlaces cableados o inalámbricos, enlaces ópticos o similares, también pueden considerarse medios que contienen el software. Como se usa en la presente memoria, a menos que
10 se limite a medios de "almacenamiento" tangibles, no transitorios, los términos tales como "medio legible" por ordenador o máquina se refieren a cualquier medio que participa en la provisión de instrucciones a un procesador para su ejecución.

15 Por lo tanto, un medio legible por máquina, tal como un código ejecutable por computadora, puede tomar muchas formas, incluyendo, pero no limitado a, un medio de almacenamiento tangible, un medio de onda portadora o un medio de transmisión física. Los medios de almacenamiento no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, tales como cualquiera de los dispositivos de almacenamiento en cualquier computadora o similares, tales como los que pueden usarse para implementar las bases de datos, etc., que se muestran en los dibujos. Los medios de almacenamiento volátiles incluyen la memoria dinámica, tal como la
20 memoria principal de dicha plataforma informática. Los medios de transmisión tangibles incluyen cables coaxiales; cable de cobre y fibra óptica, incluidos los cables que componen un bus dentro de un sistema informático. Los medios de transmisión de ondas portadoras pueden tomar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o luminosas como las generadas durante las comunicaciones de datos por radiofrecuencia (RF) e infrarrojos (IR). Las formas comunes de medios legibles por computadora incluyen,
25 por ejemplo: un disquete, un disco flexible, disco duro, cinta magnética, cualquier otro medio magnético, un CD-ROM, DVD o DVD-ROM, cualquier otro medio óptico, tarjetas perforadas, cinta de papel, cualquier otro medio físico de almacenamiento con patrones de orificios, una RAM, una ROM, una PROM y EPROM, una FLASH-EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transporta datos o instrucciones, cables o vínculos que transportan dicha onda portadora, o cualquier otro medio a partir del cual
30 una computadora puede leer códigos de programación y/o datos. Muchas de estas formas de medios legibles por ordenador pueden estar implicadas en llevar una o más secuencias de una o más instrucciones a un procesador para su ejecución.

35 El sistema informático **1901** puede incluir o estar en comunicación con una pantalla electrónica **1935** que comprende una interfaz de usuario (UI) **1940** para proporcionar, por ejemplo, datos de salida de secuencia que incluyen cromatogramas, secuencias así como bits, bytes o corrientes de bits codificadas o leídas por una máquina o sistema informático que codifica o descodifica ácidos nucleicos, datos sin procesar, archivos y archivos zip comprimidos o descomprimidos para codificarlos o descodificarlos en datos almacenados en ADN. Los ejemplos de UI incluyen, sin limitación, una interfaz gráfica de usuario (GUI) y una interfaz de usuario
40 basada en la web.

Los métodos y sistemas de la presente divulgación se pueden implementar mediante uno o más algoritmos. Se puede implementar un algoritmo por medio de software tras su ejecución por la unidad central de procesamiento **1905**. El algoritmo se puede usar, por ejemplo, con un índice de ADN y datos sin procesar o
45 datos comprimidos o descomprimidos en archivos zip, para determinar un método personalizado para codificar información digital a partir de los datos sin procesar o datos comprimidos en archivos zip, antes de codificar la información digital.

50 Ejemplos

Ejemplo 1: Codificación, escritura y lectura de un solo poema en moléculas de ADN.

Los datos a codificar son un archivo de texto que contiene un poema. Los datos se codifican manualmente con pipetas para mezclar componentes de ADN de dos capas de 96 componentes para construir identificadores
55 utilizando el esquema de producto implementado con PCR de extensión superpuesta. La primera capa, X, comprende 96 componentes de ADN en total. La segunda capa, Y, también comprende 96 componentes en total. Antes de escribir el ADN, los datos se mapean a binario y luego se recodifican en un formato de peso uniforme donde cada cadena contigua (disjunta adyacente) de 61 bits de los datos originales se traduce a una cadena de 96 bits con exactamente 17 valores de bits de 1. Este formato de peso uniforme puede tener
60 cualidades naturales de verificación de errores. Luego, los datos se combinan en una tabla de 96 por 96 para formar un mapa de referencia.

65 El panel medio de la FIG. 18A muestra el mapa de referencia bidimensional de una tabla de 96 por 96 que codifica el poema en una pluralidad de identificadores. Los puntos oscuros corresponden a un valor de bit '1' y los puntos blancos corresponden a un valor de bit '0'. Los datos se codifican en identificadores utilizando dos capas de 96 componentes. A cada valor X y valor Y de la tabla se le asigna un componente y los componentes

X e Y se ensamblan en un identificador utilizando PCR de extensión superpuesta para cada coordenada (X, Y) con un valor '1'. Los datos se volvieron a leer (p. ej., se descodificaron) secuenciando la biblioteca de identificadores para determinar la presencia o ausencia de cada posible ensamblaje (X,Y).

- 5 El panel derecho de la figura FIG. 18A muestra un mapa de calor bidimensional de las abundancias de secuencias presentes en la biblioteca de identificadores según se determina mediante secuenciación. Cada píxel representa una molécula que comprende los componentes X e Y correspondientes, y la intensidad de la escala de grises en ese píxel representa la abundancia relativa de esa molécula en comparación con otras moléculas. Los identificadores se toman como los 17 ensamblajes (X, Y) más abundantes en cada fila (ya que la codificación de peso uniforme garantiza que cada cadena contigua de 96 bits pueda tener exactamente 17 valores '1' y, por lo tanto, 17 identificadores correspondientes).

Ejemplo 2: Codificación de un archivo de texto de 62.824 bits.

- 15 Los datos a codificar son un archivo de texto de tres poemas con un total de 62.824 bits. Los datos se codifican utilizando un Manipulador de Líquidos Labcyte Echo® para mezclar componentes de ADN de dos capas de 384 componentes para construir identificadores utilizando el esquema de producto implementado con PCR de extensión superpuesta. La primera capa, X, comprende 384 componentes de ADN en total. La segunda capa, Y, también comprende 384 componentes en total. Antes de escribir el ADN, los datos se mapean a binario y
- 20 luego se recodifican para disminuir el peso (número de valores de bits de '1') e incluir sumas de comprobación. Las sumas de comprobación se establecen de manera que exista un identificador que corresponda a una suma de comprobación para cada cadena contigua de 192 bits de datos. Los datos recodificados tienen un peso de aproximadamente 10.100, que corresponde al número de identificadores que se va a construir. Luego, los datos se pueden dividir en una tabla de 384 por 384 para formar un mapa de referencia.
- 25 El panel medio de la FIG. 18B muestra un mapa de referencia bidimensional de una tabla de 384 por 384 que codifica el archivo de texto en una pluralidad de identificadores. Cada coordenada (X,Y) corresponde al bit de datos en la posición $X + (Y-1) * 192$. Los puntos negros corresponden a un valor de bit de '1' y los puntos blancos corresponden a un valor de bit de '0'. Los puntos negros en el lado derecho de la figura son las sumas de comprobación y el patrón de puntos negros en la parte superior de la figura es el libro de códigos (p. ej., un diccionario para descodificar los datos). A cada valor X y valor Y de la tabla se le pueden asignar ejemplos proporcionados en la memoria descriptiva. Si bien la invención se ha descrito con referencia a la memoria descriptiva antes mencionada, no se pretende que las descripciones e ilustraciones de las realizaciones de la presente memoria se interpreten en un sentido limitante. A los expertos en la técnica se les ocurrirán numerosas variaciones, cambios y sustituciones sin alejarse de la invención. Además, se entenderá que todos los aspectos de la invención no se limitan a las representaciones, configuraciones o proporciones relativas específicas establecidas en la presente memoria que dependen de una variedad de condiciones y variables. Se entenderá que, en la práctica de la invención, se pueden emplear varias alternativas a las realizaciones de la invención descritas en la presente memoria. Por lo tanto, se contempla que la invención también cubra dichas alternativas, modificaciones o variaciones. El alcance de protección de la presente invención está definido por las reivindicaciones 1-24 adjuntas.
- 30
- 35
- 40

REIVINDICACIONES

1. Un método para escribir información en secuencias de ácido nucleico, comprendiendo el método:

5 (a) traducir dicha información en una cadena de símbolos, en donde cada símbolo en dicha cadena de símbolos tiene uno de dos valores de símbolo posibles;

(b) codificar dicha cadena de símbolos en una pluralidad de identificadores, en donde un identificador individual de dicha pluralidad de identificadores comprende una pluralidad correspondiente de componentes, en donde
10 cada componente en dicha pluralidad de componentes comprende una secuencia de ácido nucleico distintiva con un primer y un segundo extremo, una primera región de hibridación en el primer extremo de la secuencia de ácido nucleico distintiva y una segunda región de hibridación en el segundo extremo de la secuencia de ácido nucleico distintiva, en donde la pluralidad de componentes se dividen en M capas, en donde cada capa de dichas M capas comprende un conjunto distintivo de componentes y en donde cada identificador individual de
15 dicha pluralidad de identificadores corresponde a un símbolo individual de dicha cadena de símbolos;

(c) construir identificadores individuales de dicha pluralidad de identificadores depositando la correspondiente pluralidad de componentes en un compartimento y ensamblando, mediante una o más reacciones, un componente de cada una de las M capas en un orden fijo; y

20 (d) construir una biblioteca de identificadores que comprende dichos identificadores individuales correspondientes a dicha cadena de símbolos, en donde el valor del símbolo en cada posición de dicha cadena de símbolos está representado por la presencia o ausencia de un identificador distintivo en la biblioteca de identificadores.

25 2. El método de la reivindicación 1, en donde los dos valores de símbolo posibles son valores de bit de 0 y 1, en donde el símbolo individual con el valor de bit de 0 en la cadena de símbolos está representado por la ausencia de un identificador distintivo en la biblioteca de identificadores, en donde el símbolo individual con el valor de bit de 1 en la cadena de símbolos está representado por la presencia de un identificador distintivo en
30 la biblioteca de identificadores.

3. El método de la reivindicación 1, en donde dentro de cada capa, los componentes tienen una primera región de hibridación común y una segunda región de hibridación común.

35 4. El método de la reivindicación 3, en donde los componentes en capas adyacentes tienen regiones de hibridación complementarias que permiten que los componentes interaccionen entre sí.

5. El método de la reivindicación 4, en donde los componentes en una primera capa y en una segunda capa tienen regiones de hibridación complementarias de manera que cualquier componente de la primera capa es
40 capaz de unirse a cualquier componente de la segunda capa.

6. El método de la reivindicación 5, en donde los componentes en la segunda capa y en una tercera capa tienen regiones de hibridación complementarias de manera que cualquier componente de la segunda capa es capaz de unirse a cualquier componente de la primera capa en un extremo, y es capaz de unirse a cualquier
45 componente de la tercera capa en un extremo opuesto.

7. El método de la reivindicación 1, en donde el ensamblaje de la pluralidad de componentes entre sí comprende usar reacción en cadena de la polimerasa (PCR) de extensión superpuesta, ensamblaje de ciclado de polimerasa, ligación de extremos cohesivos, ensamblaje de bioladrillos, ensamblaje golden gate, ensamblaje de Gibson, ensamblaje de recombinasa, reacción de ciclado de ligasa o ligación dirigida por molde.
50

8. El método de la reivindicación 1, en donde se combinan una o más bibliotecas de identificadores y en donde cada biblioteca de identificadores de dicha una o más bibliotecas de identificadores está etiquetada con un código de barras distintivo.
55

9. El método de la reivindicación 1, en donde dichos identificadores individuales comprenden nucleótidos reactivos configurados para la unión de un resto químico, siendo el resto químico al menos uno de fluorescente, quimioluminiscente, ácido, básico, hidrófobo, hidrófilo, oxidativo o reactivo.

60 10. El método de la reivindicación 1, en donde cada uno de dichos identificadores individuales comprende un código de barras distintivo que tiene una distancia de Hamming mínima con respecto a los códigos de barras de otros identificadores individuales.

11. El método de la reivindicación 1, en donde se construye combinatoriamente un objeto de rango a partir de componentes en (b), y un objeto de valor de byte identifica un valor de byte y se construye combinatoriamente a partir de componentes en (b).
65

12. El método de la reivindicación 11, en donde el objeto de rango identifica una posición relativa de un byte en una corriente de bits.
- 5 13. El método de la reivindicación 1, en donde la biblioteca de identificadores comprende una pluralidad de secuencias de ácido nucleico, y en donde dicha pluralidad de secuencias de ácido nucleico almacena metadatos de dicha información y/u oculta dicha información.
- 10 14. El método de la reivindicación 13, en donde dichos metadatos comprenden información secundaria correspondiente a una fuente de dicha información, un destinatario previsto de dicha información, un formato original de dicha información, instrumentación y métodos utilizados para codificar dicha información, una fecha y una hora de escritura de dicha información en dicha biblioteca de identificadores, modificaciones realizadas a dicha información y/o una referencia a otra información.
- 15 15. El método de la reivindicación 1, que comprende además extraer un subconjunto diana de dicha pluralidad de identificadores de dicha biblioteca de identificadores.
- 20 16. El método de la reivindicación 15, que comprende además combinar una pluralidad de sondas con dicha biblioteca de identificadores, en donde dicha pluralidad de sondas comparte complementariedad con dicho uno o más componentes de dicho identificador individual de dicho subconjunto diana de dicha biblioteca de identificadores.
- 25 17. El método de la reivindicación 16, en donde dicha pluralidad de sondas comprende una o más etiquetas de afinidad y en donde dicha una o más etiquetas de afinidad es capturada por una perla de afinidad o una columna de afinidad.
- 30 18. El método de la reivindicación 15, en donde dicho identificador individual comprende una o más regiones de unión a cebadores comunes, una o más regiones de unión a cebadores variables o cualquier combinación de las mismas.
- 35 19. El método de la reivindicación 18, que comprende además combinar dicha biblioteca de identificadores con cebadores que se unen a dicha una o más regiones de unión a cebadores comunes o a dicha una o más regiones de unión a cebadores variables.
- 40 20. El método de la reivindicación 19, que comprende además unir dichos cebadores a dicha una o más regiones de unión de cebadores variables para amplificar selectivamente dicho subconjunto diana de dicha biblioteca de identificadores.
21. El método de la reivindicación 15, que comprende además eliminar selectivamente una parte de identificadores no diana de dicha biblioteca de identificadores.
22. El método de la reivindicación 21, en donde la eliminación se realiza mediante escisión selectiva con nucleasa.
- 45 23. El método de la reivindicación 22, que comprende además combinar dicha biblioteca de identificadores con Cas9 y sondas guía y en donde dichas sondas guía, guían a dicha Cas9 para eliminar identificadores específicos de dicha biblioteca de identificadores.
- 50 24. El método de la reivindicación 22, en donde dichos identificadores individuales son monocatenarios, y que comprende además combinar dicha biblioteca de identificadores con una o unas endonucleasas específicas de cadena única.

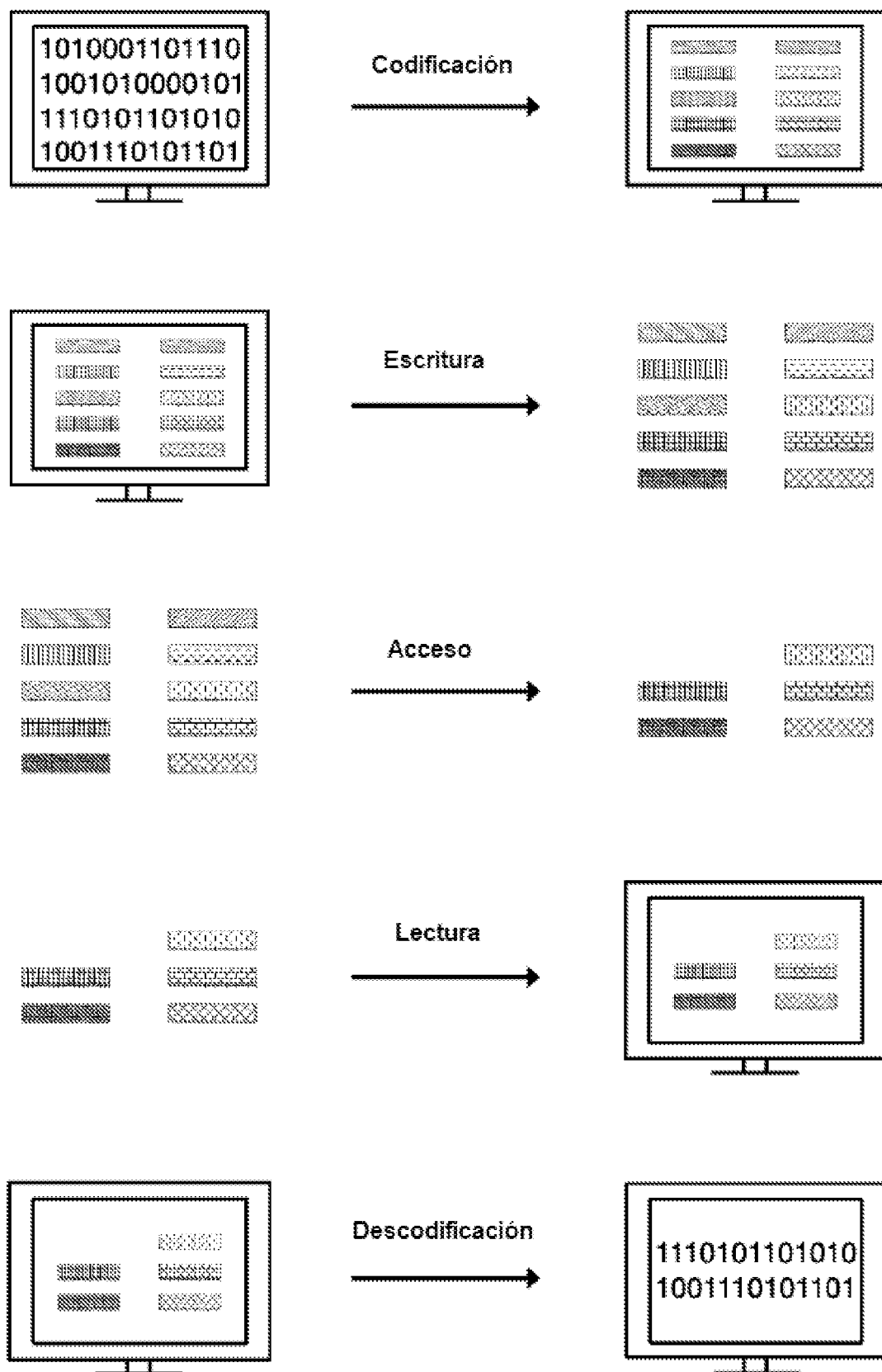


FIG. 1

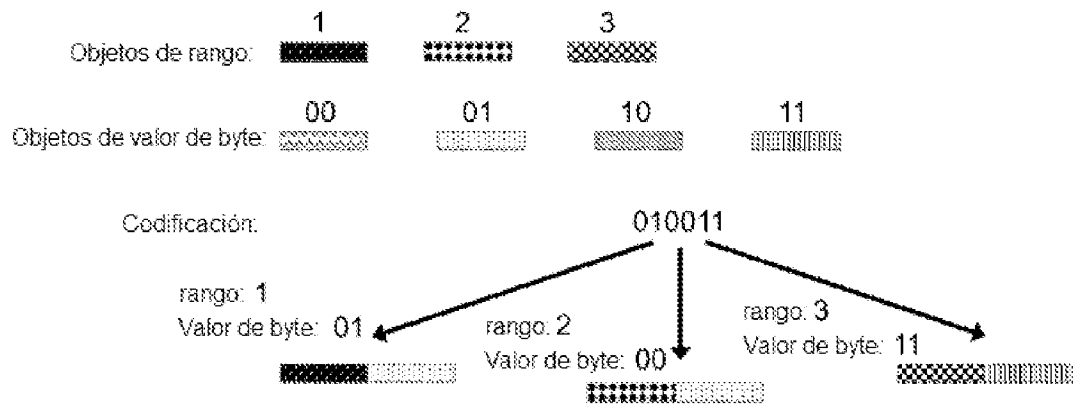


FIG. 2A

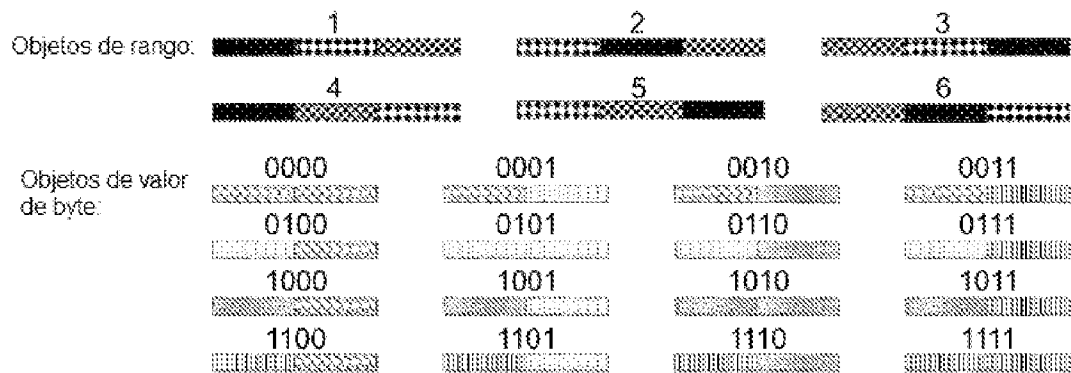


FIG. 2B

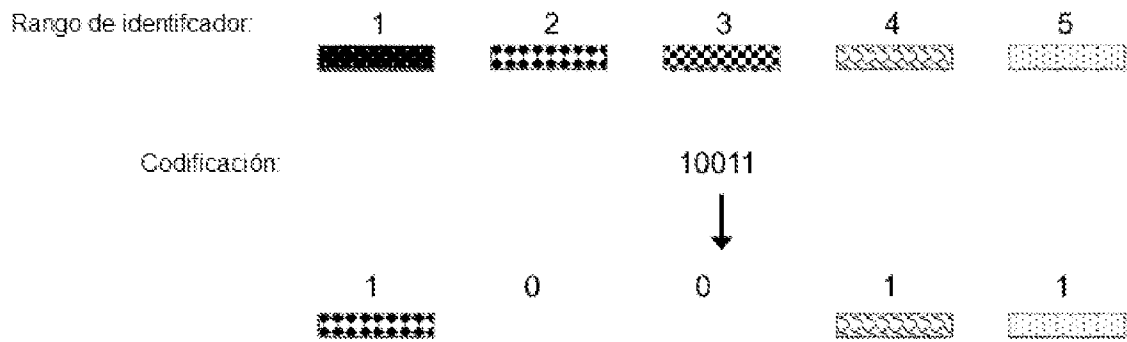


FIG. 3A

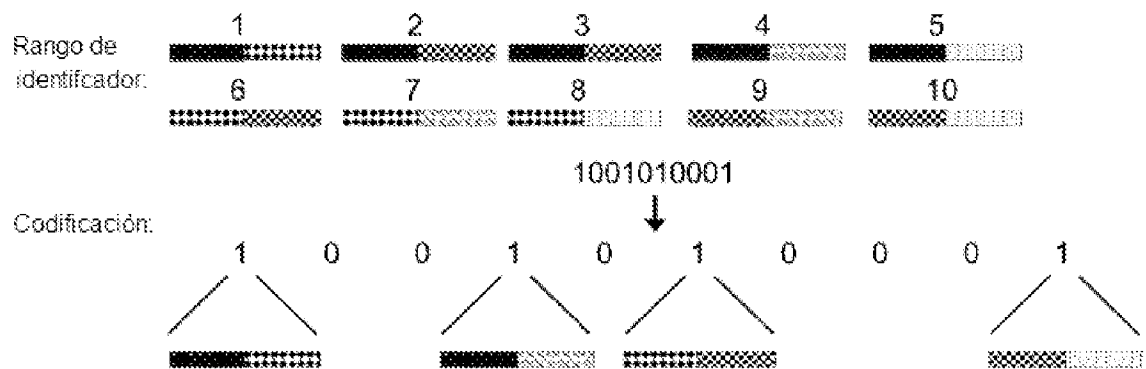


FIG. 3B

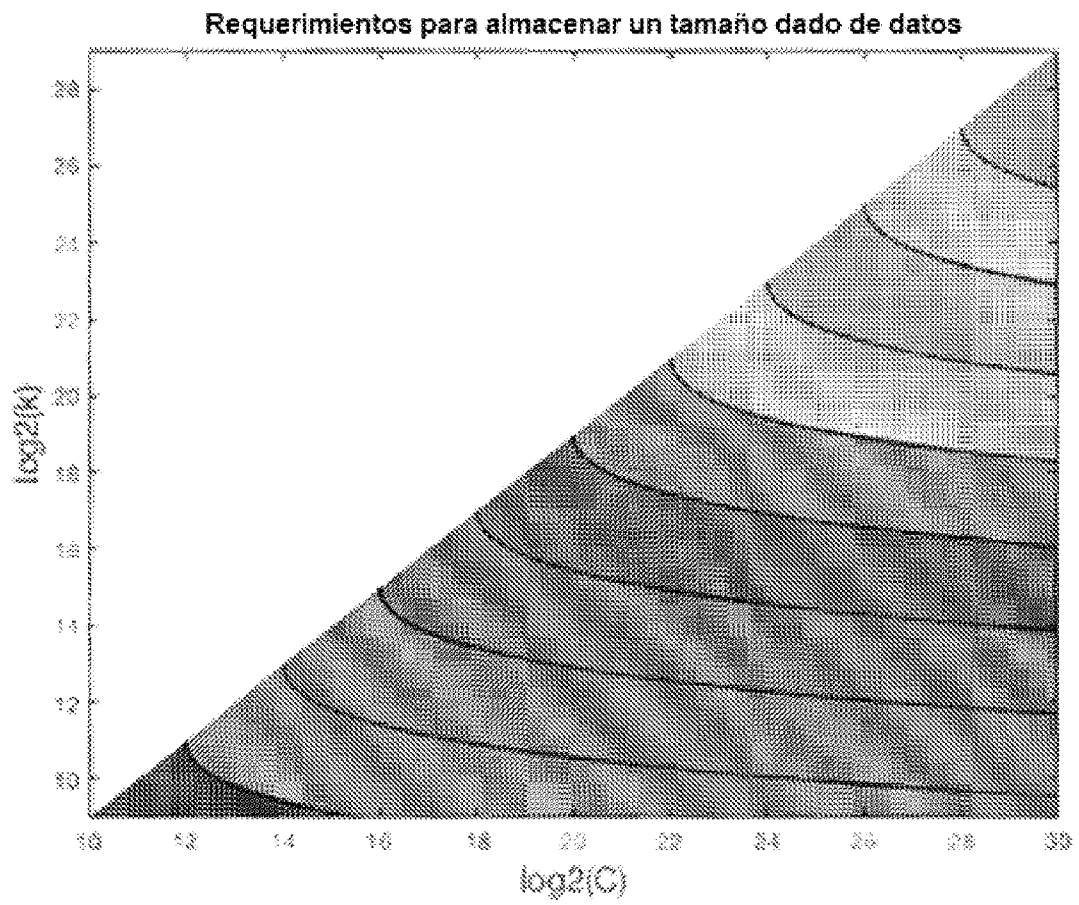


FIG. 4

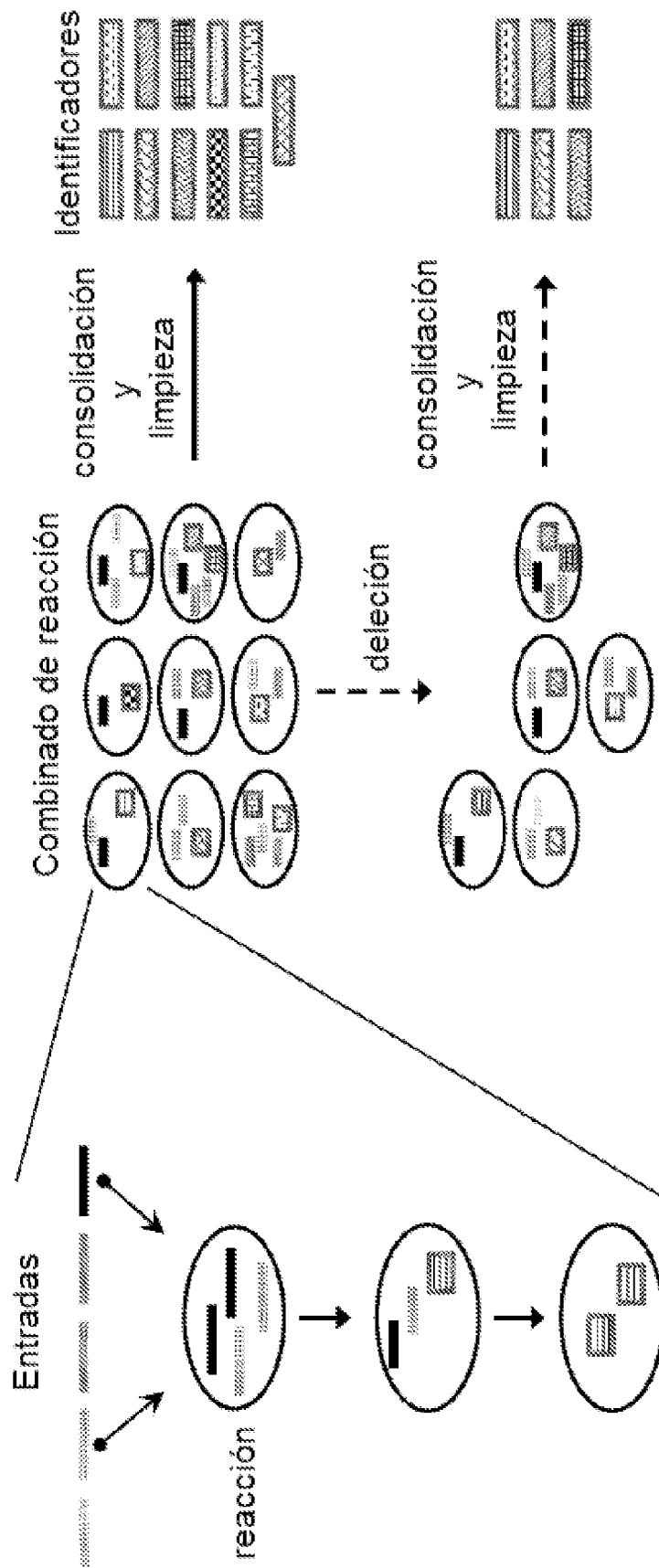
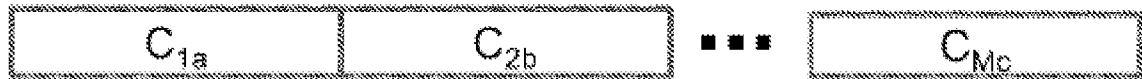


FIG. 5

C_{xy} es el y-ésimo componente en la capa x. Para una biblioteca de partida con M capas, cada una con N componentes, los identificadores tendrían la siguiente arquitectura:



donde a, b, c representan elementos en el conjunto $\{1, 2, \dots, N\}$

FIG. 6A

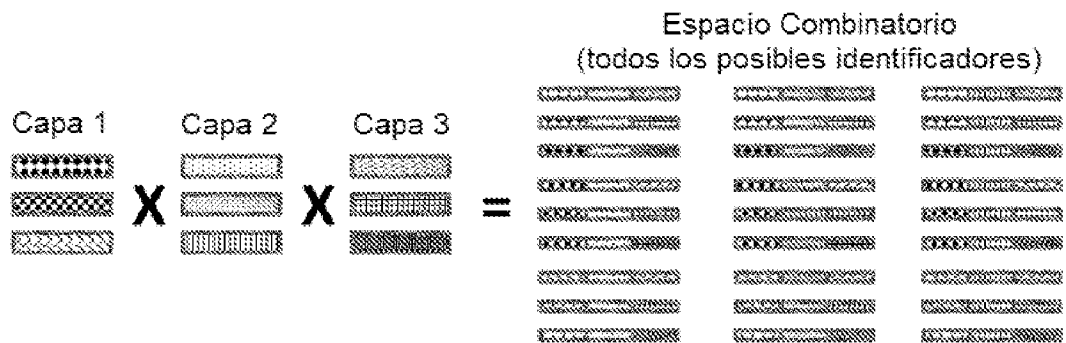


FIG. 6B

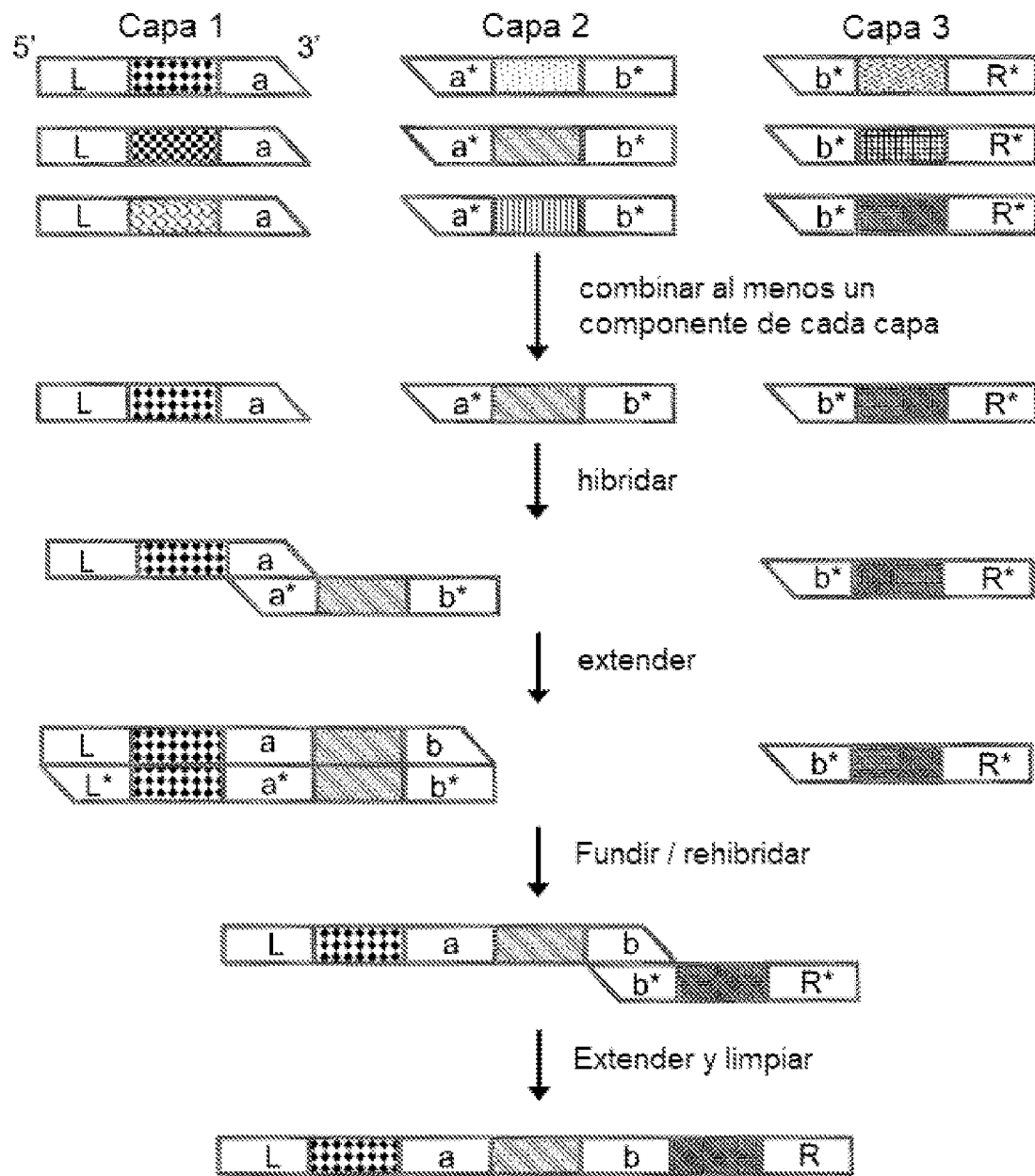


FIG. 7

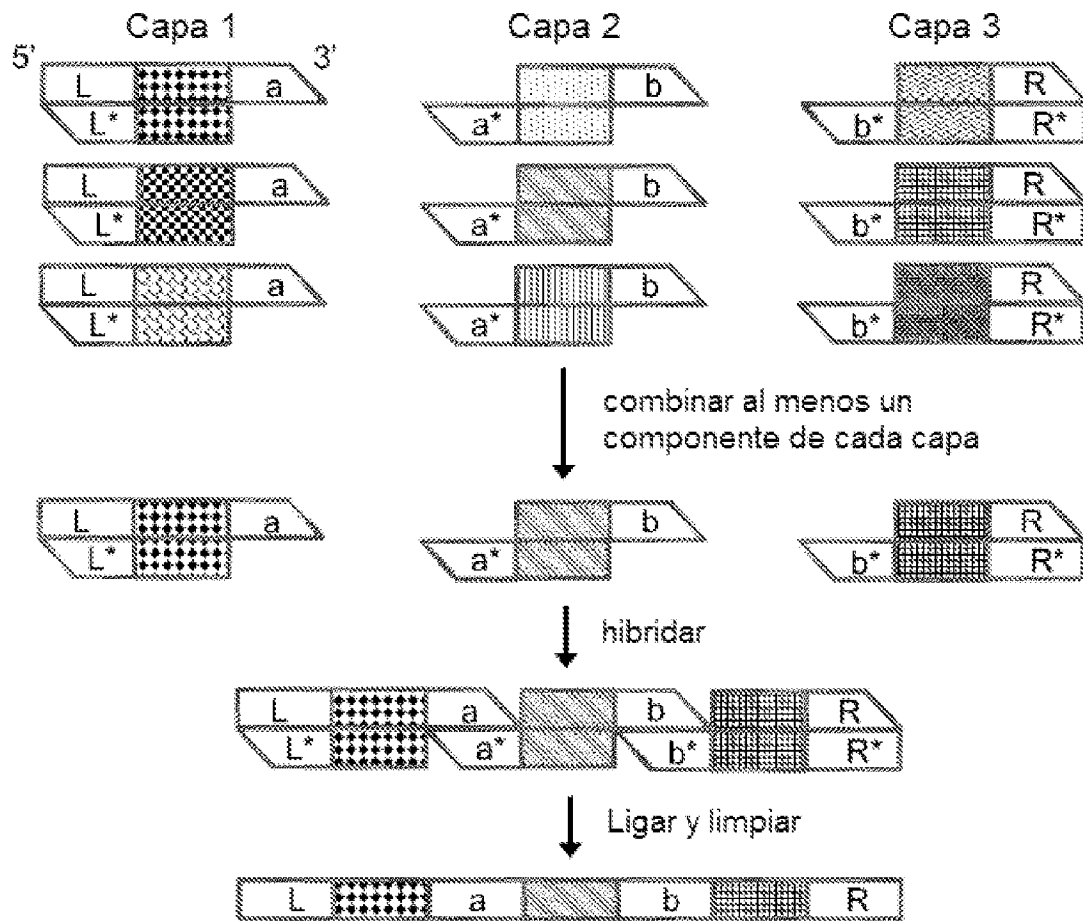


FIG. 8

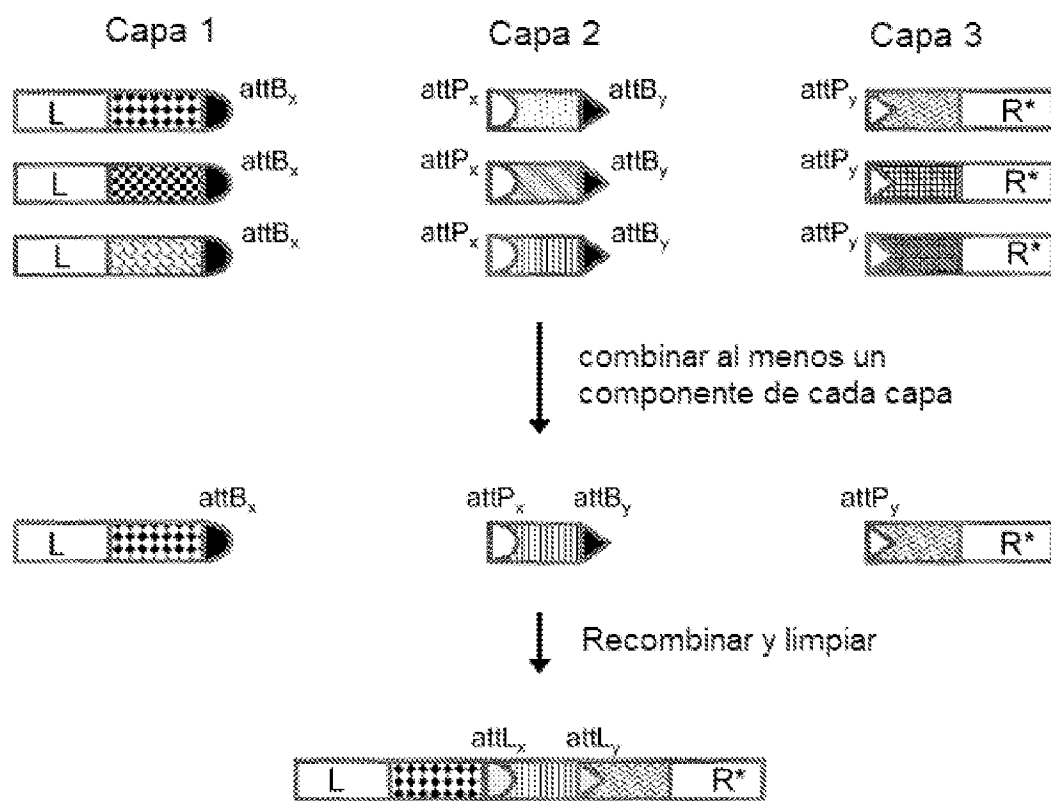
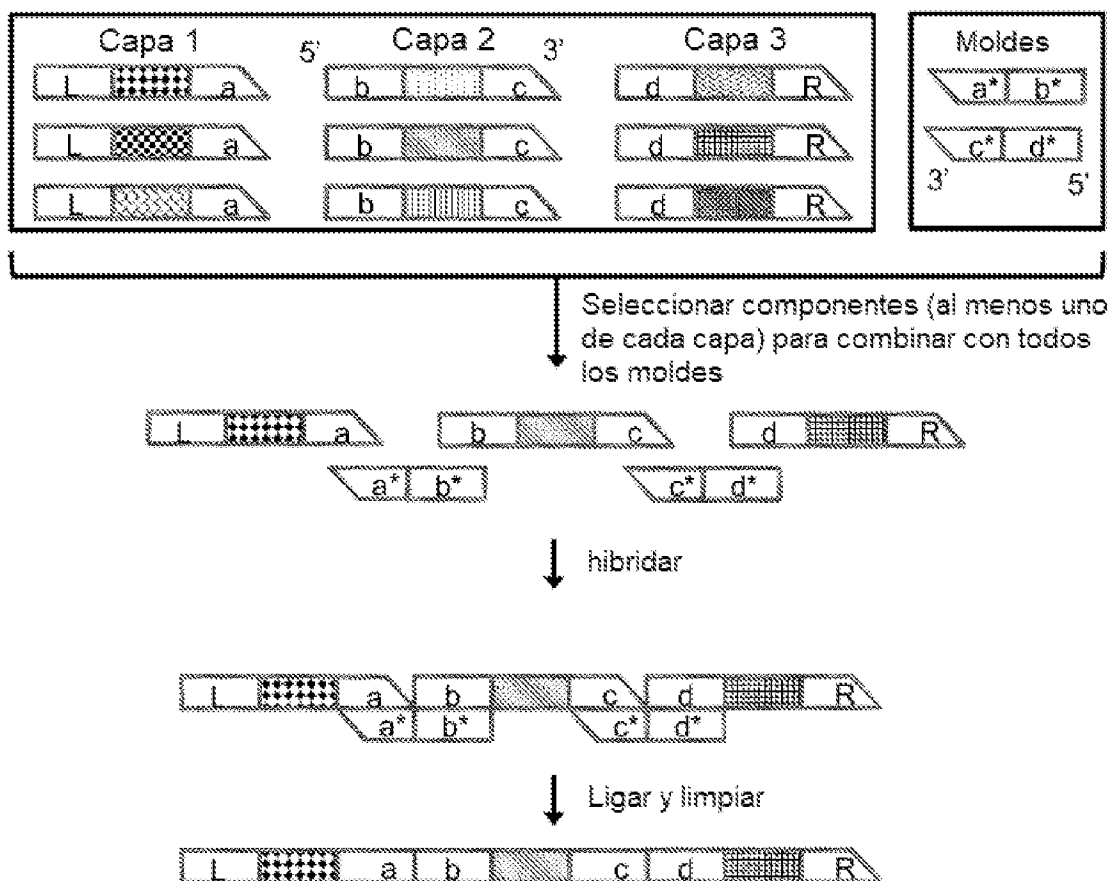


FIG. 9

**FIG. 10A**

Número de copias de 256 secuencias de ADN distintivas
ensambladas con TDL de 6 capas

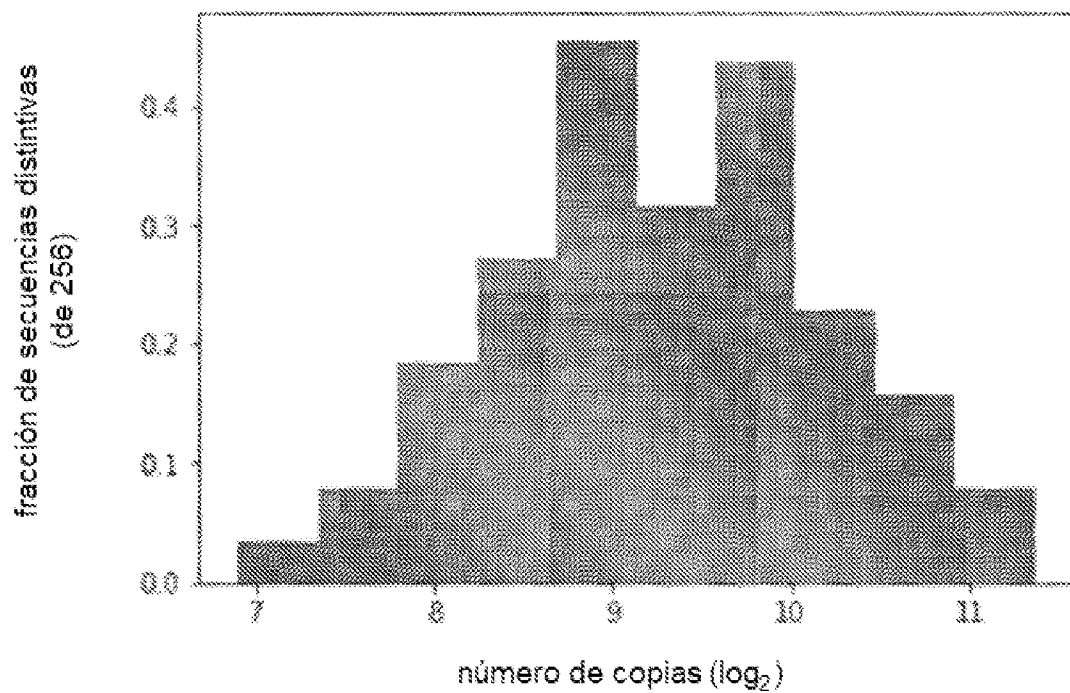


FIG. 10B

C_{xy} es el y -ésimo componente en la capa x . Para una biblioteca de partida con M capas, cada una con N componentes, los identificadores tendrían la siguiente arquitectura:



donde b, e, h son elementos en el conjunto $\{1, 2, \dots, N\}$ y a, d, f son elementos únicos en el conjunto $\{1, 2, \dots, M\}$.

El espacio combinatorio de posibles identificadores es $N^M M!$

FIG. 11A

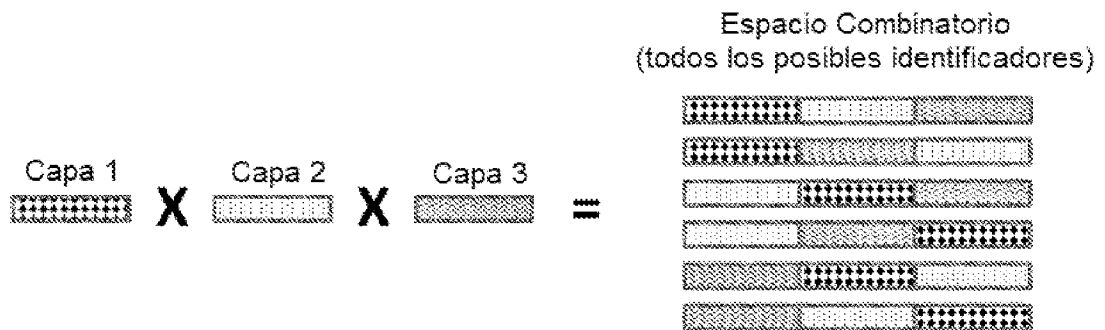


FIG. 11B

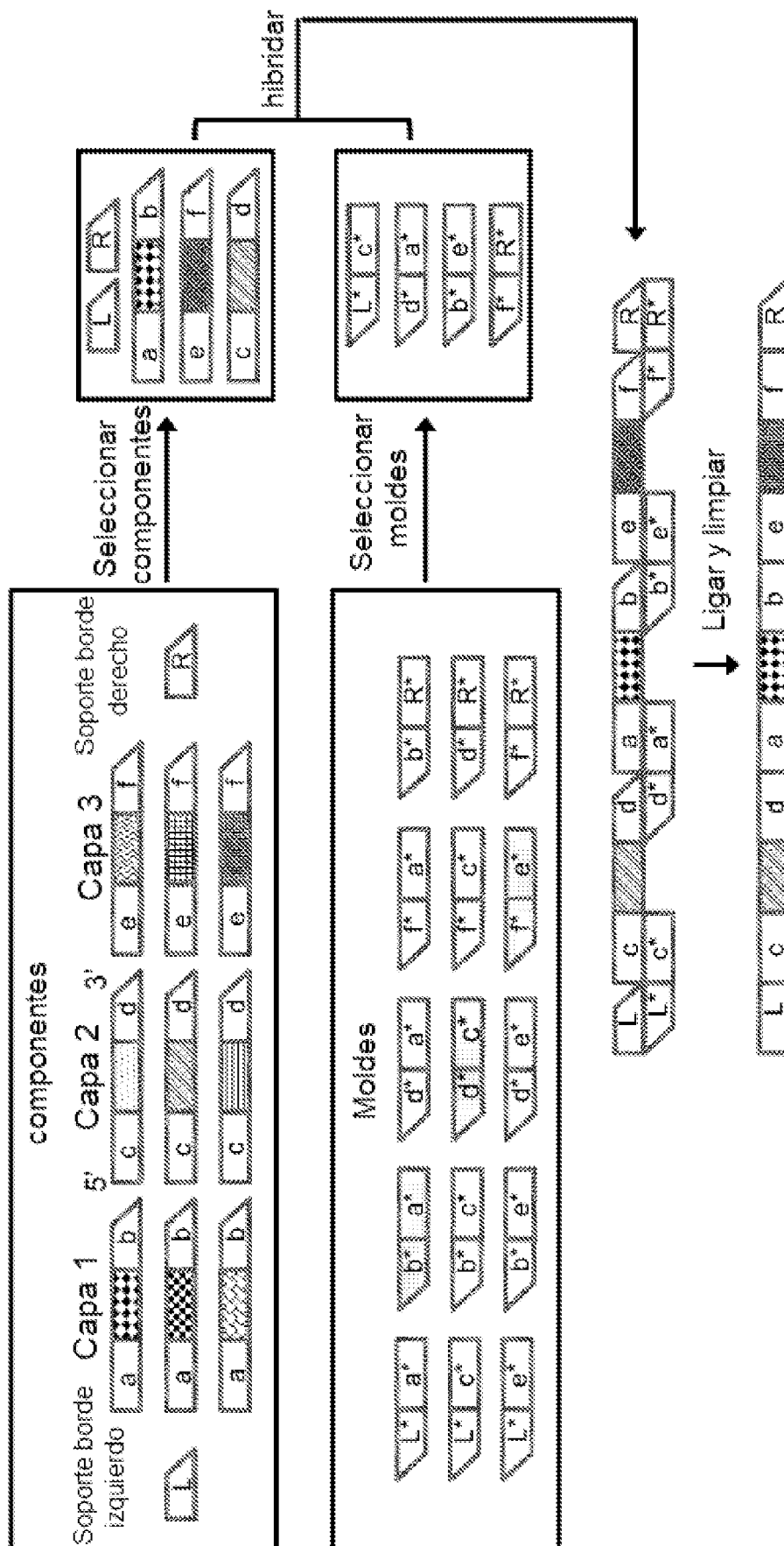


FIG. 11C

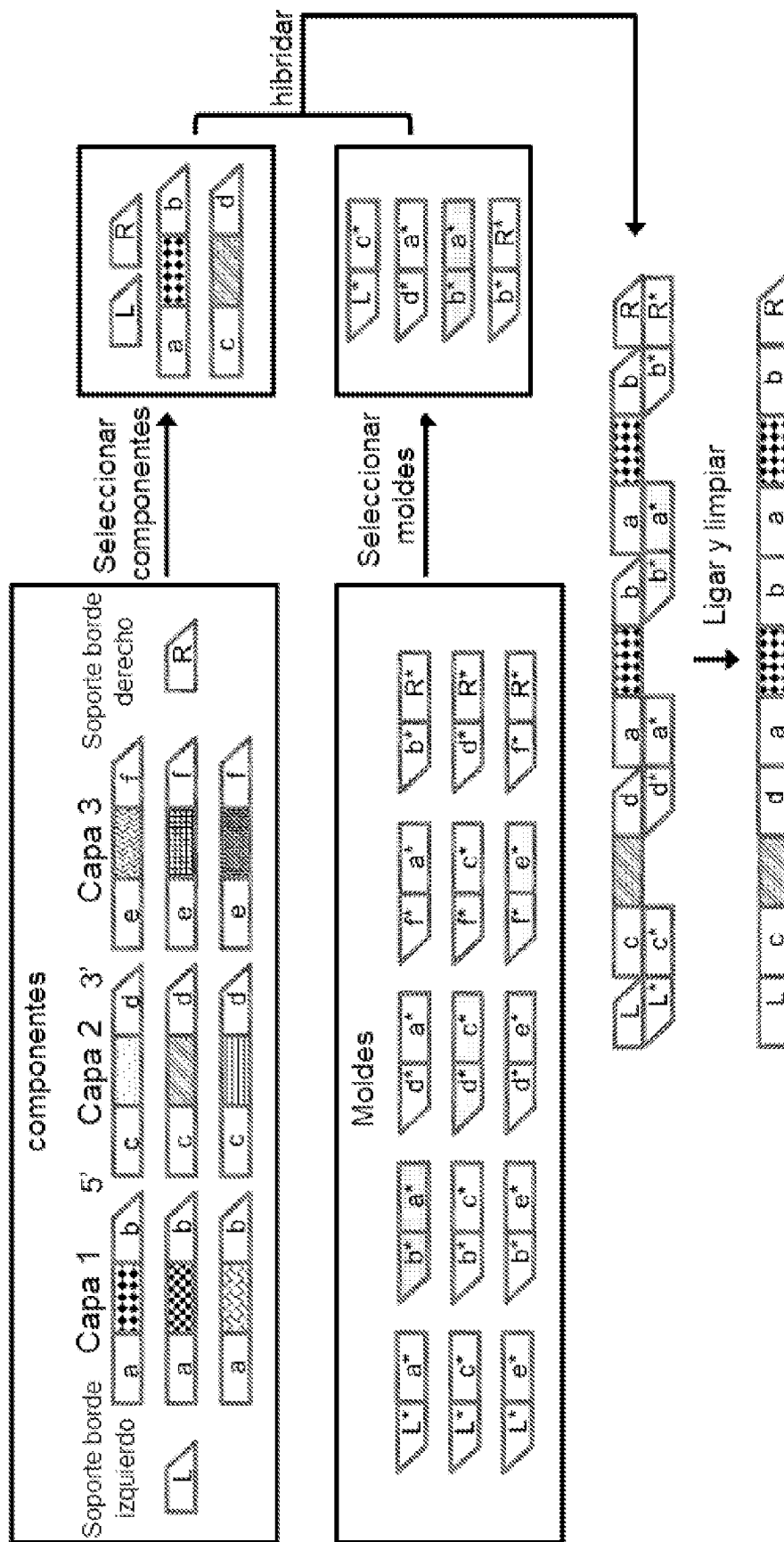


FIG. 11D

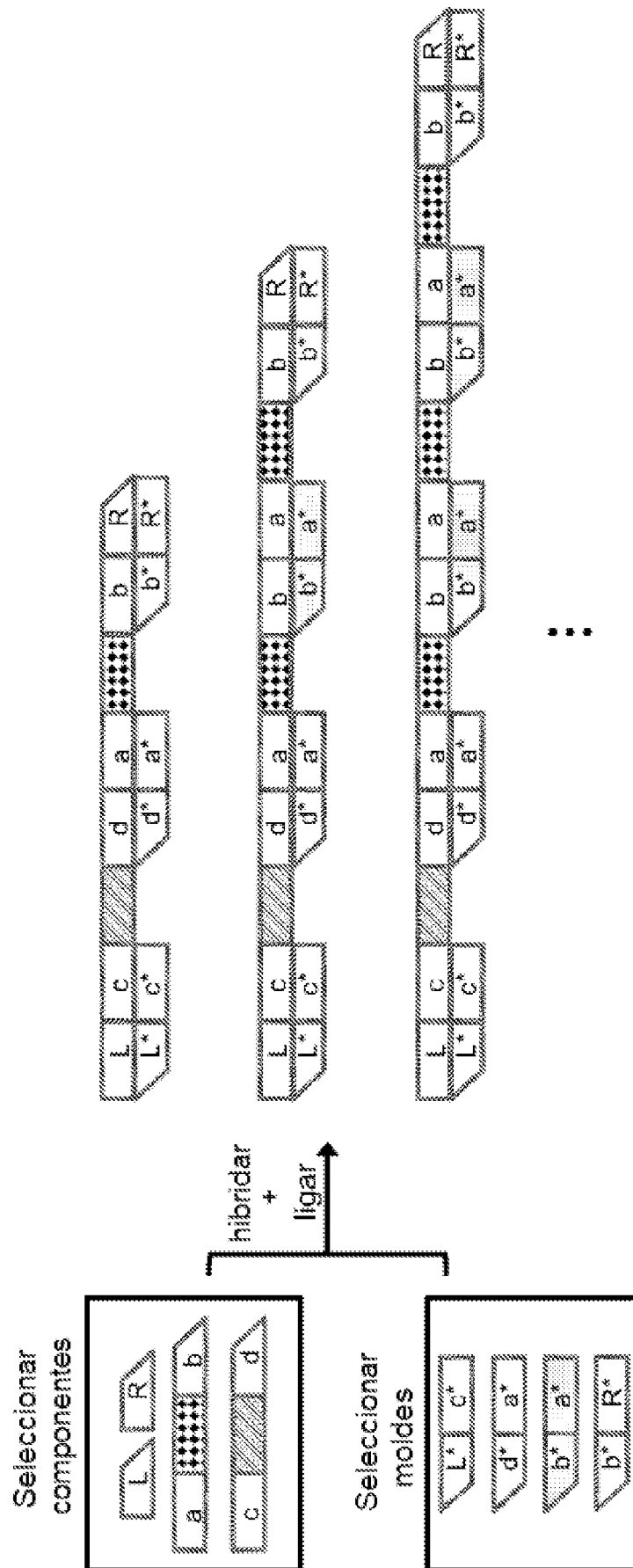


FIG. 11E

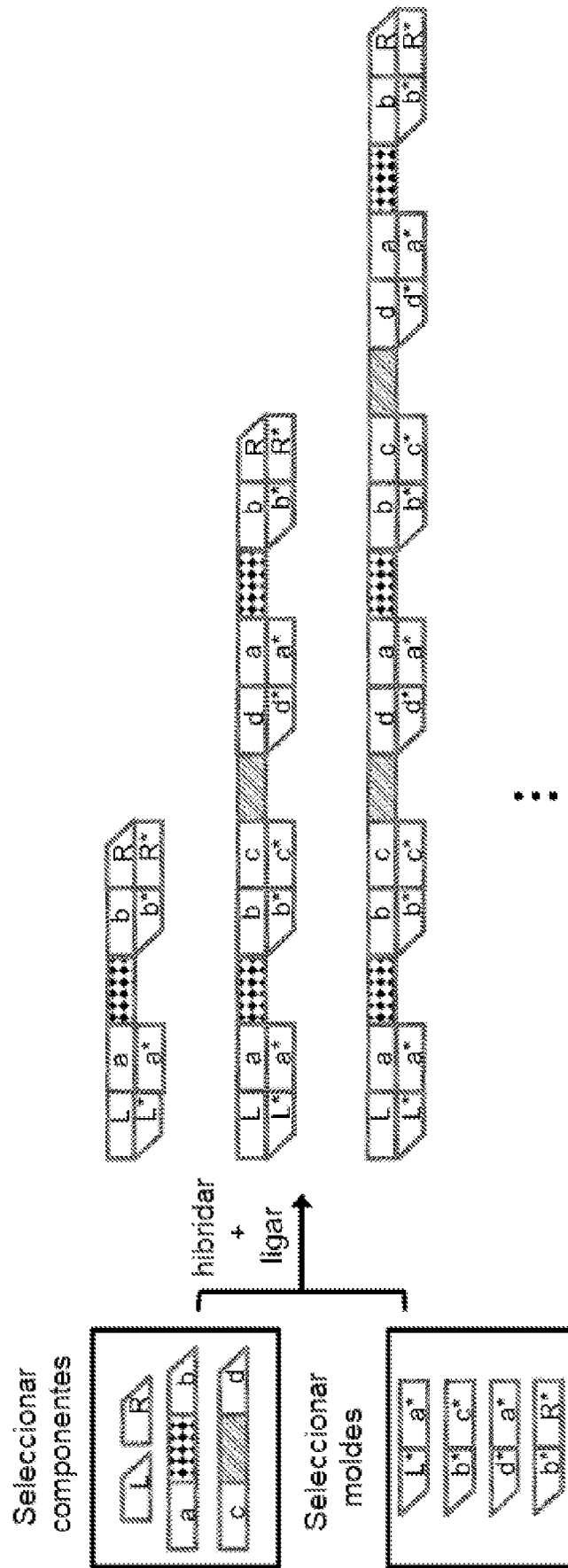


FIG. 11F

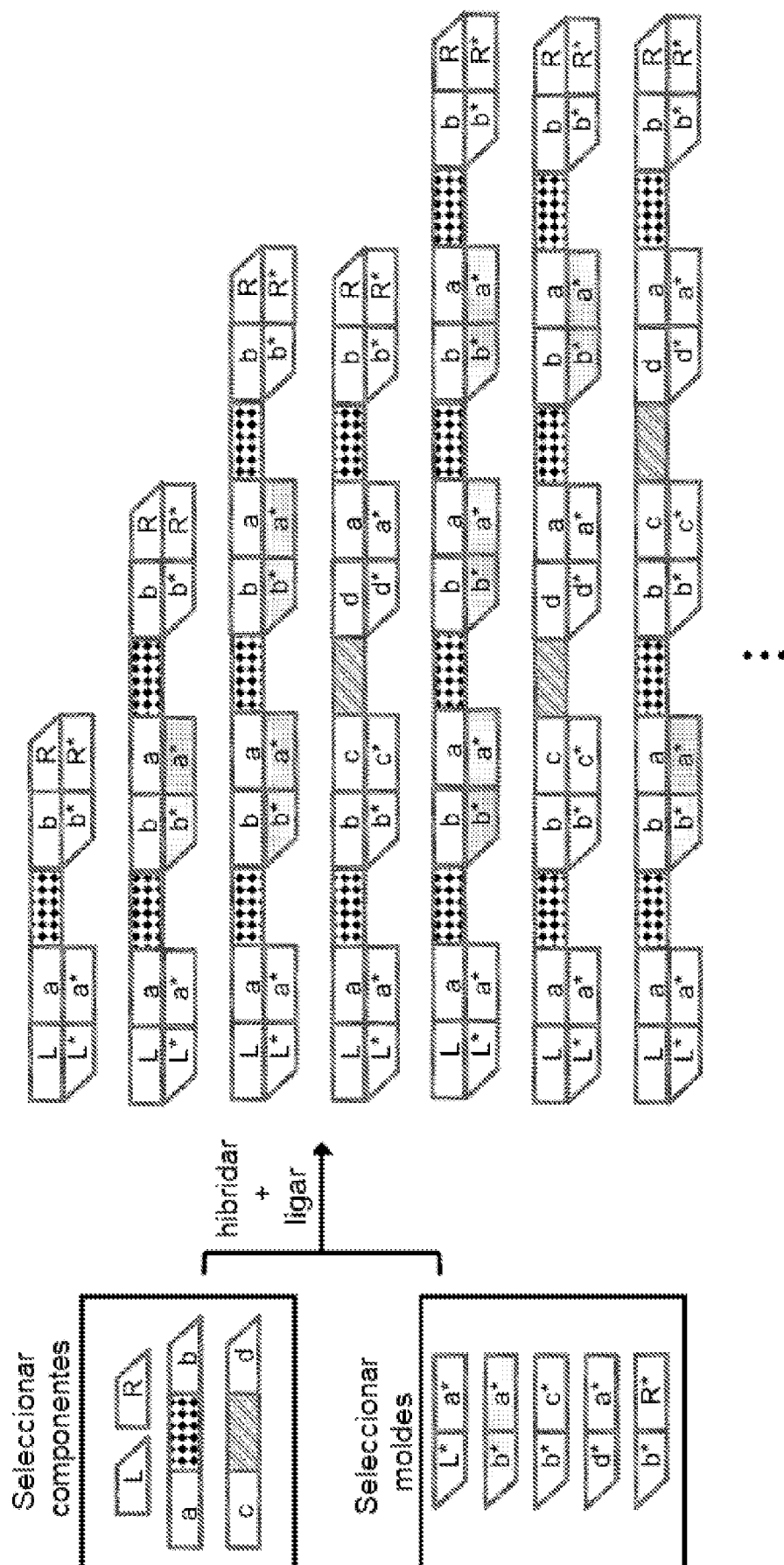


FIG. 11G

C_{xy} es el y -ésimo componente en la capa x . Para una biblioteca de partida con M capas, cada una con N componentes, los Identificadores ensamblan un número fijo de componentes, k , y tienen la siguiente arquitectura:



donde b, e, h son elementos en el conjunto $\{1, 2, \dots, N\}$ y a, d, f entre k elementos únicos, ordenados por rango en el conjunto $\{1, 2, \dots, M\}$.

El espacio combinatorio de posibles identificadores es $N^k M \text{choose } k$

FIG. 12A

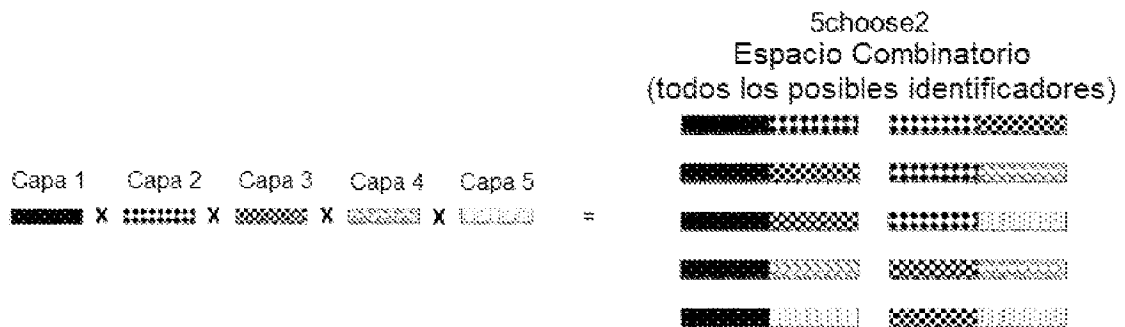


FIG. 12B

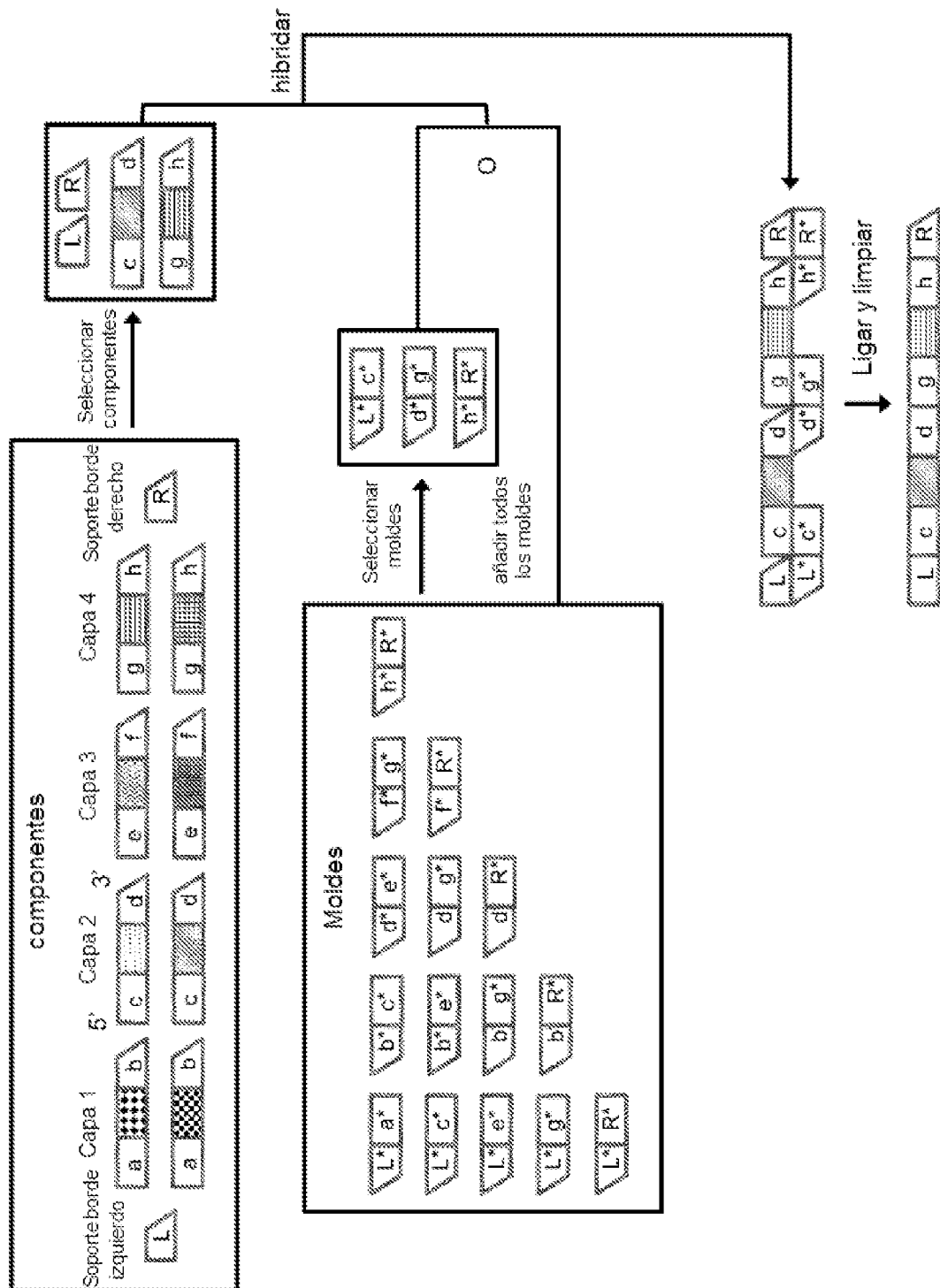
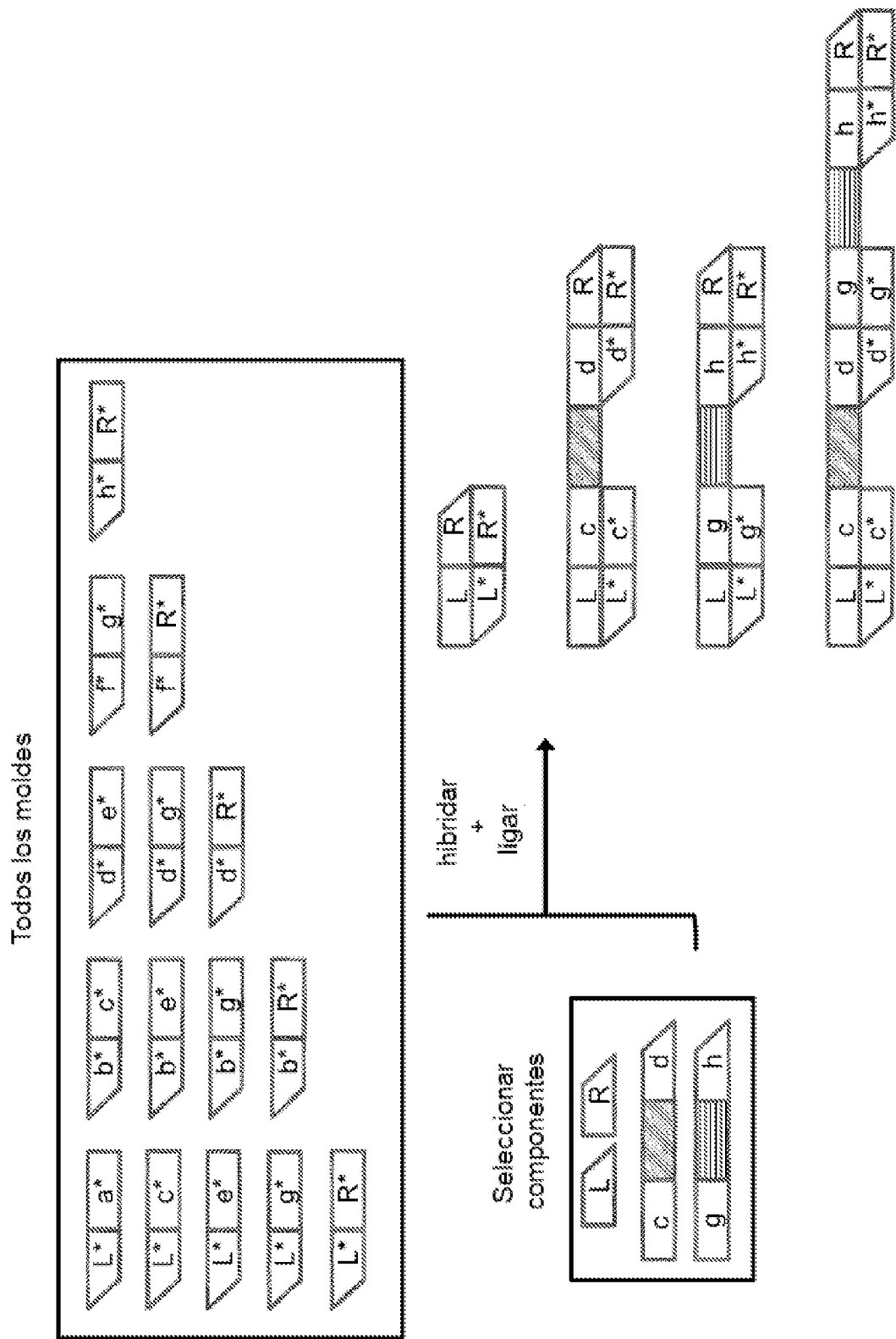


FIG. 12C



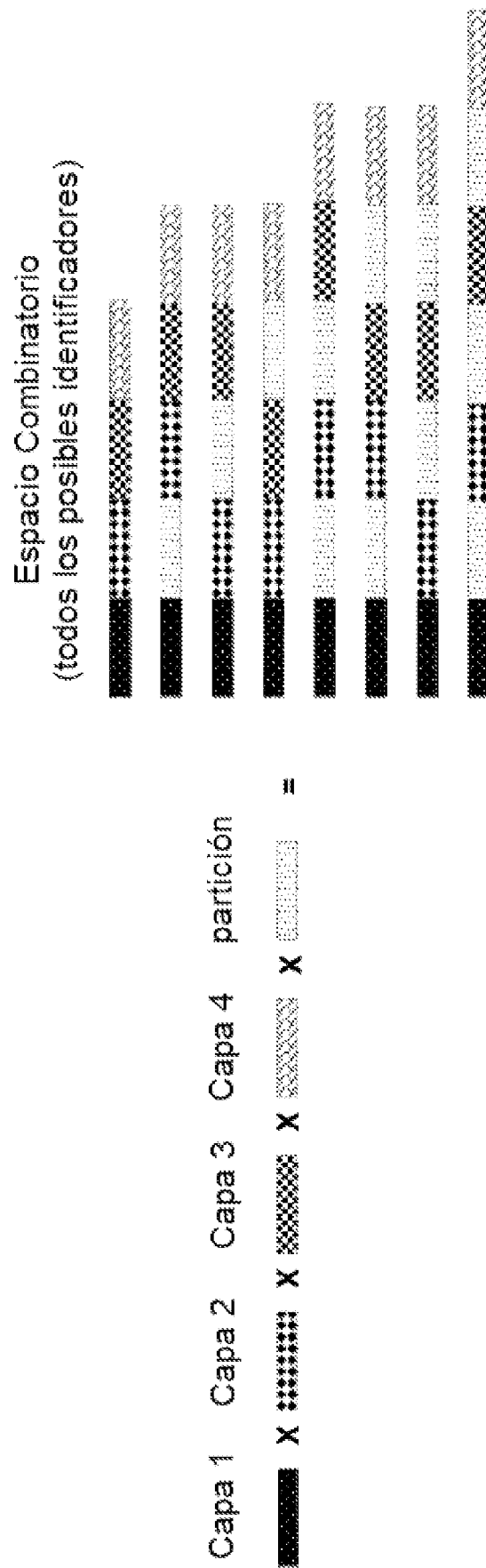


FIG. 13A

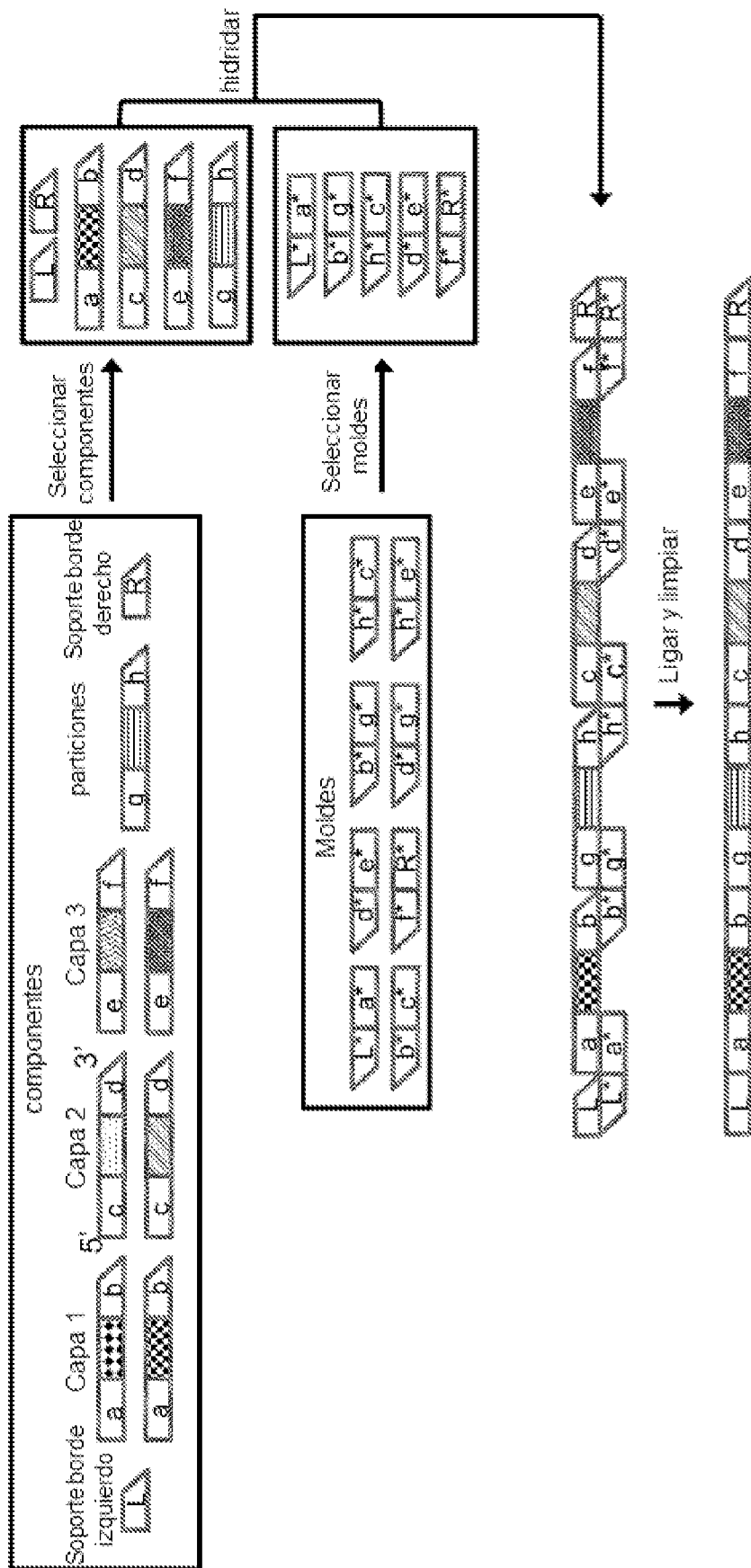


FIG. 13B

longitud de 3 componentes
Espacio Combinatorio
(todos los posibles identificadores)

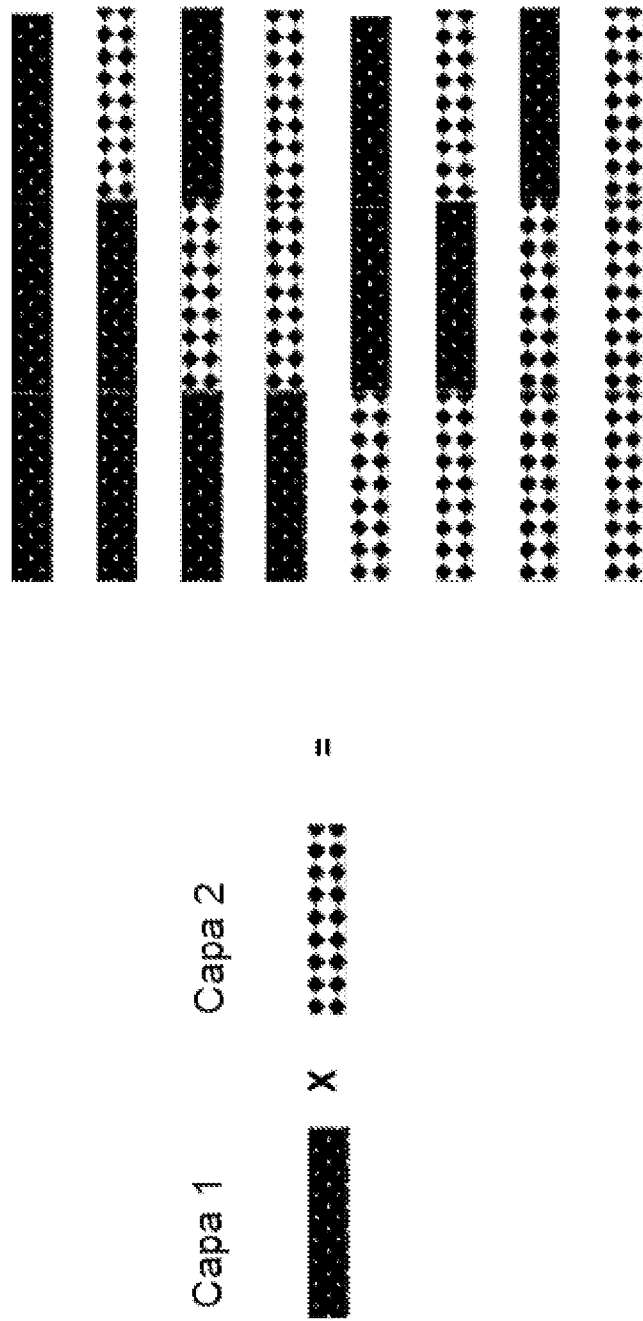


FIG. 14A

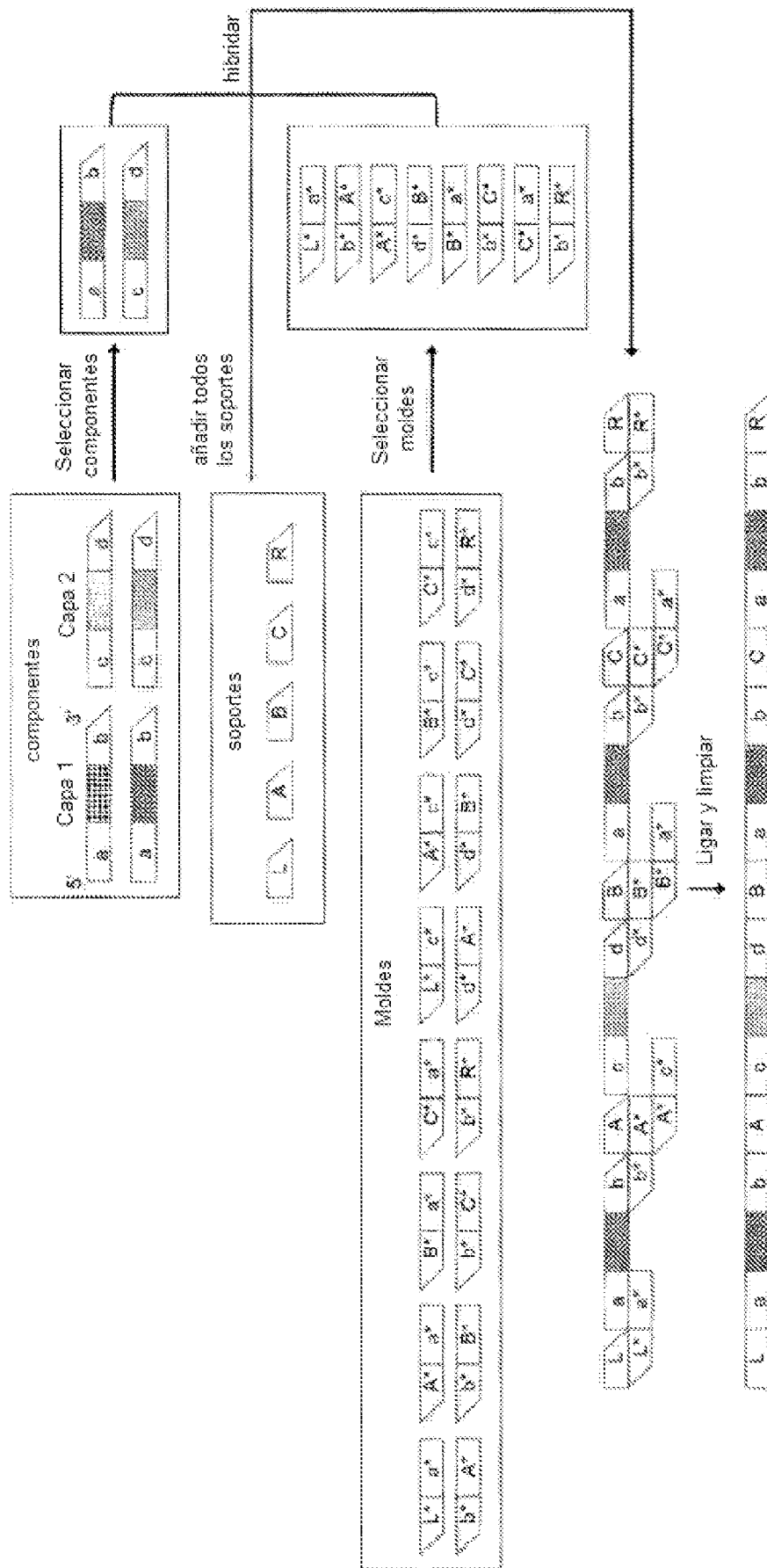


FIG. 14B

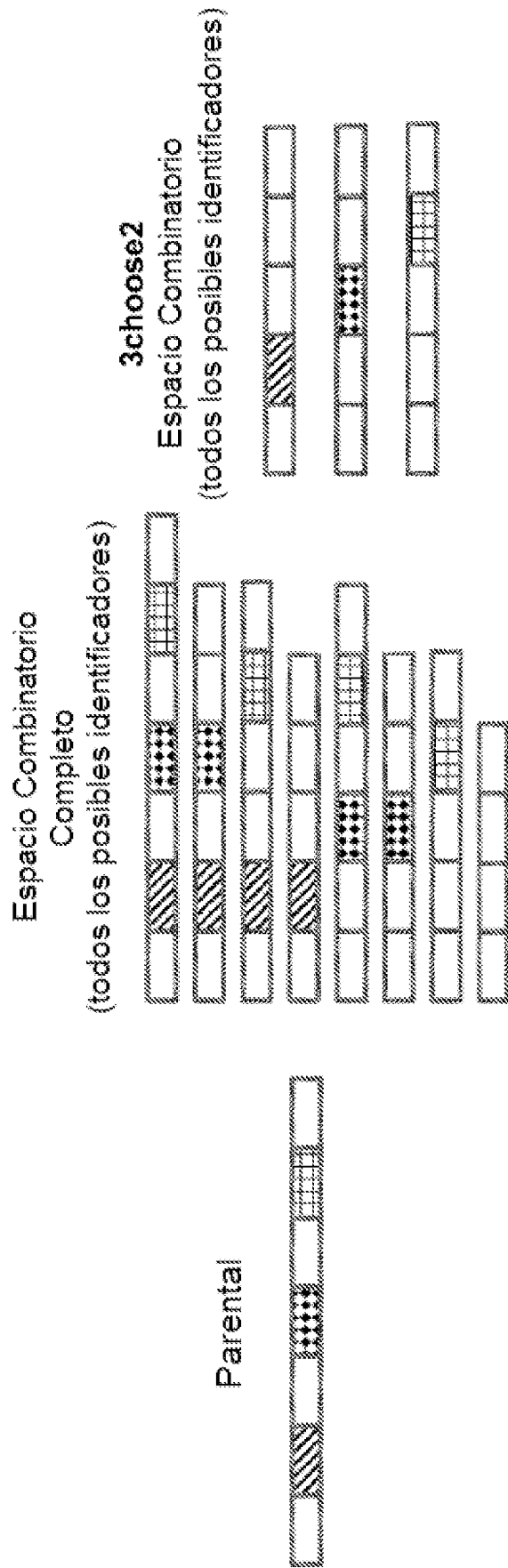


FIG. 15A

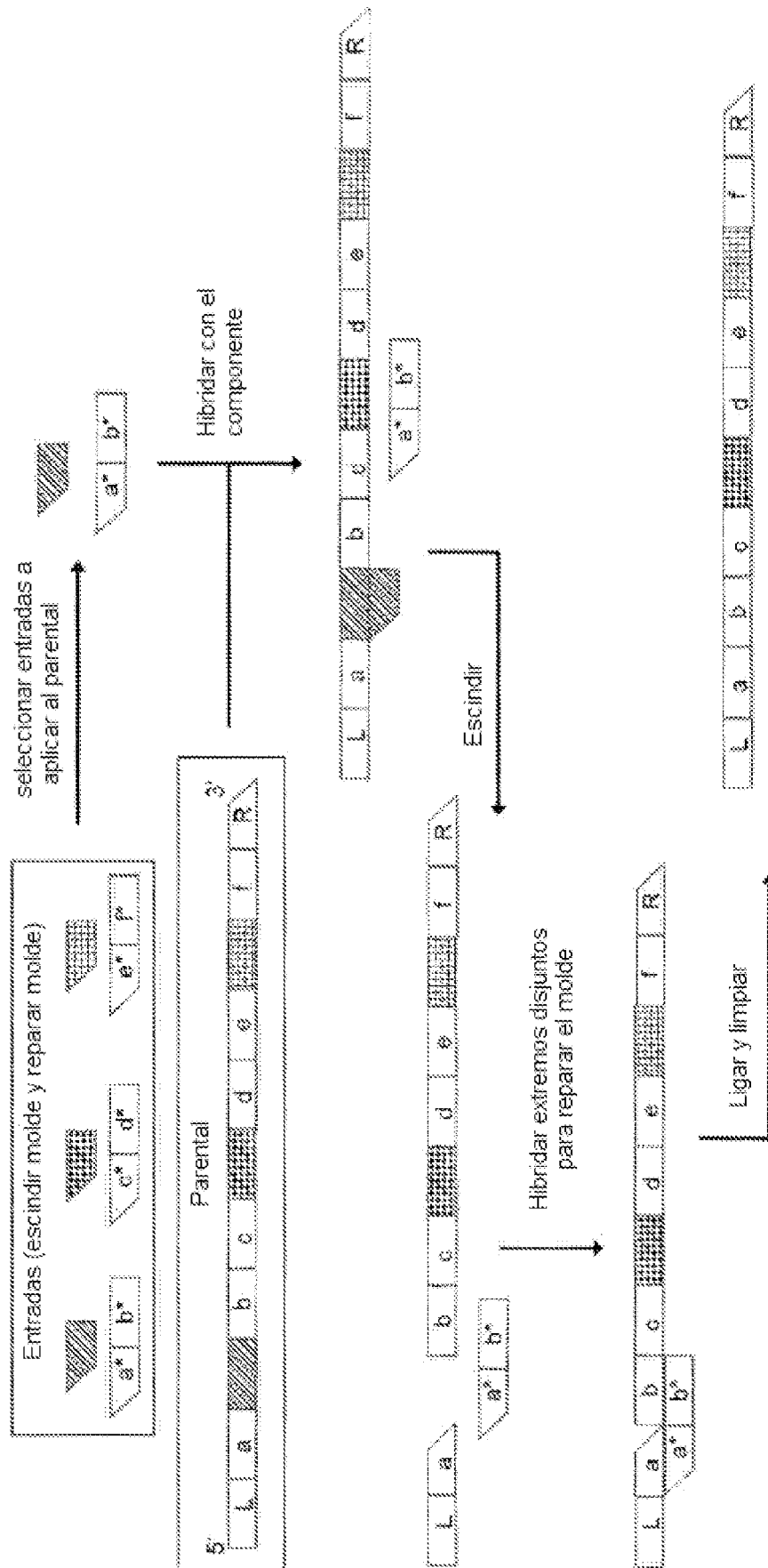


FIG. 15B



FIG. 16

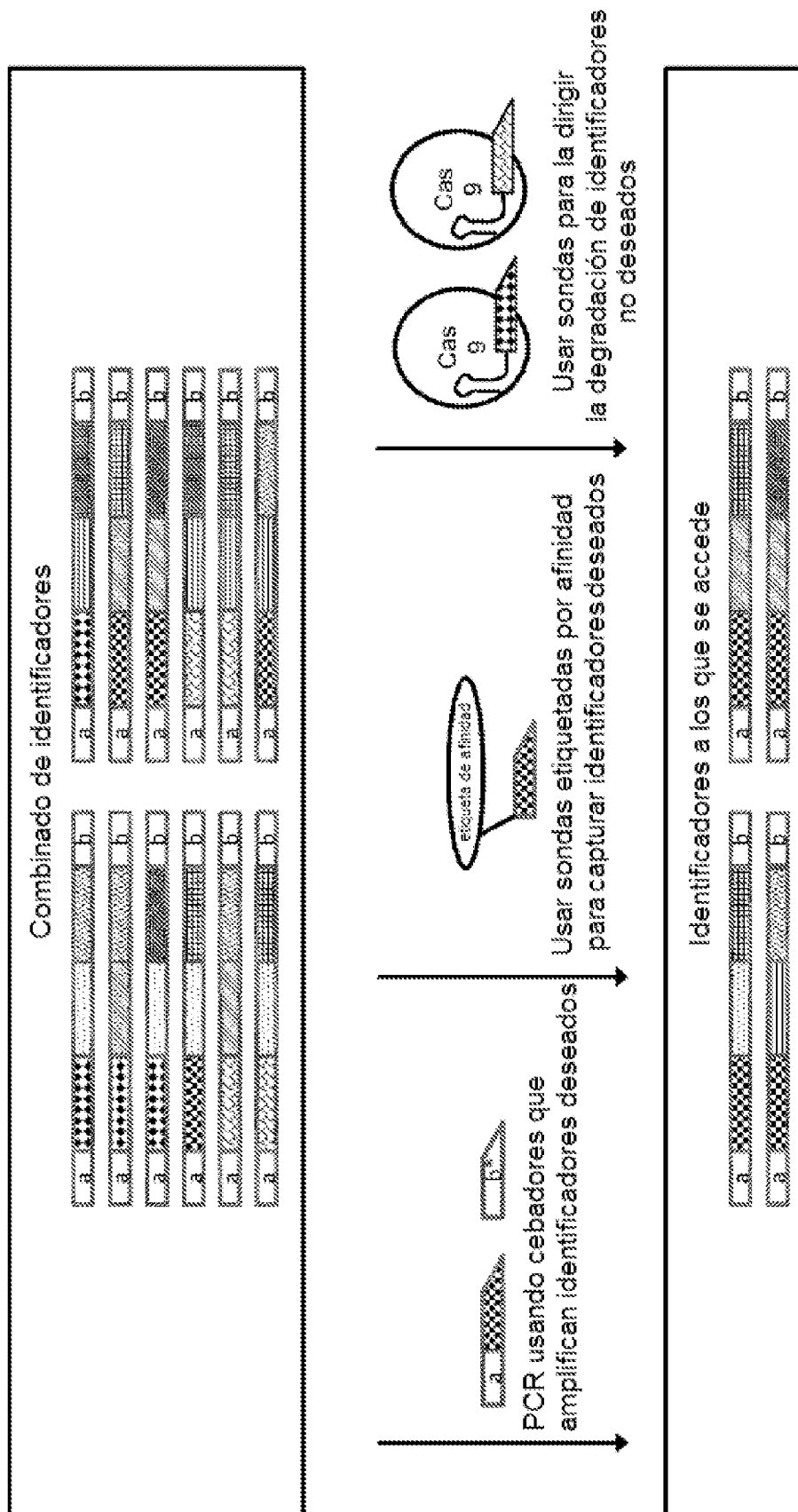


FIG. 17A

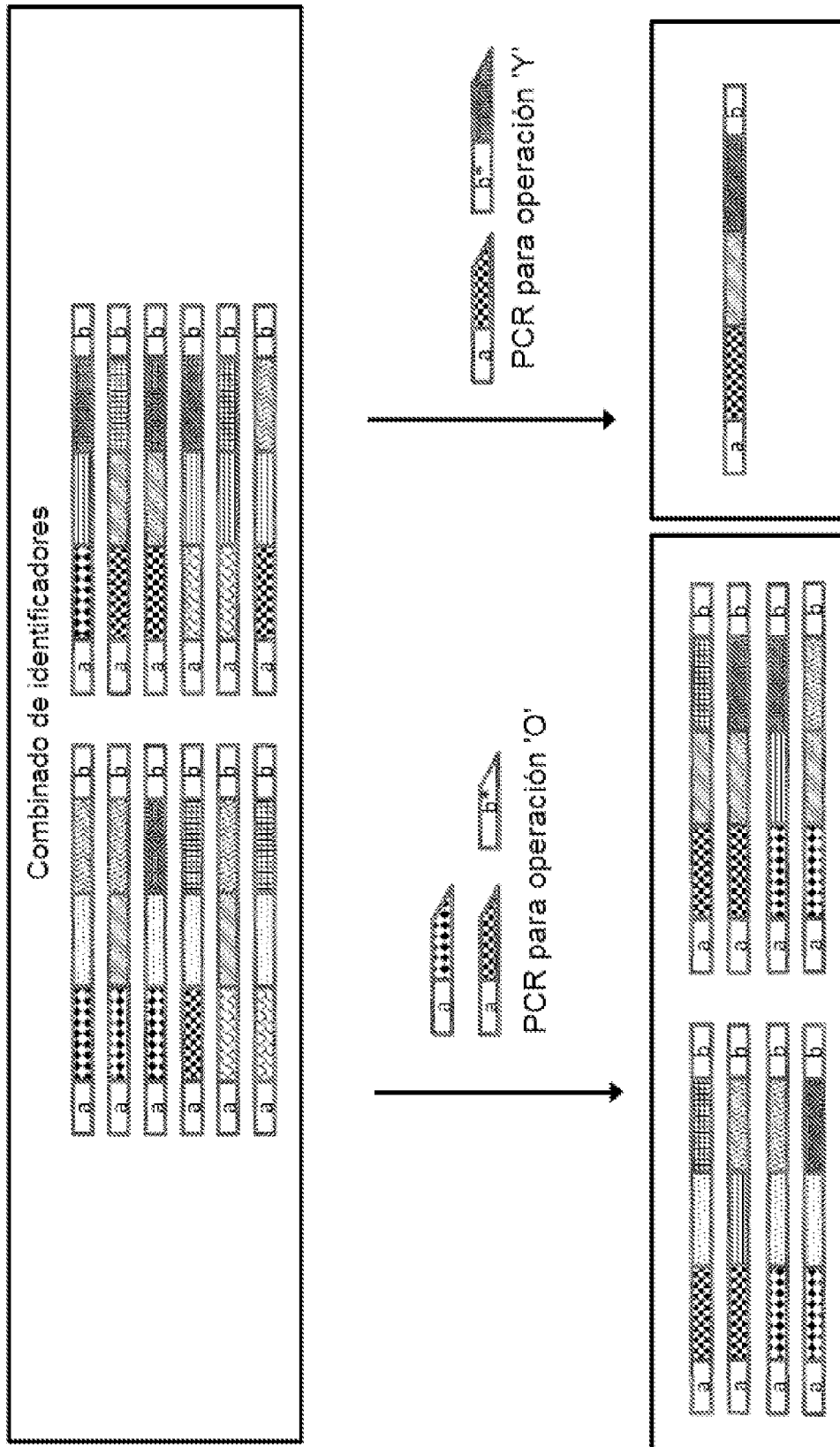


FIG. 17B

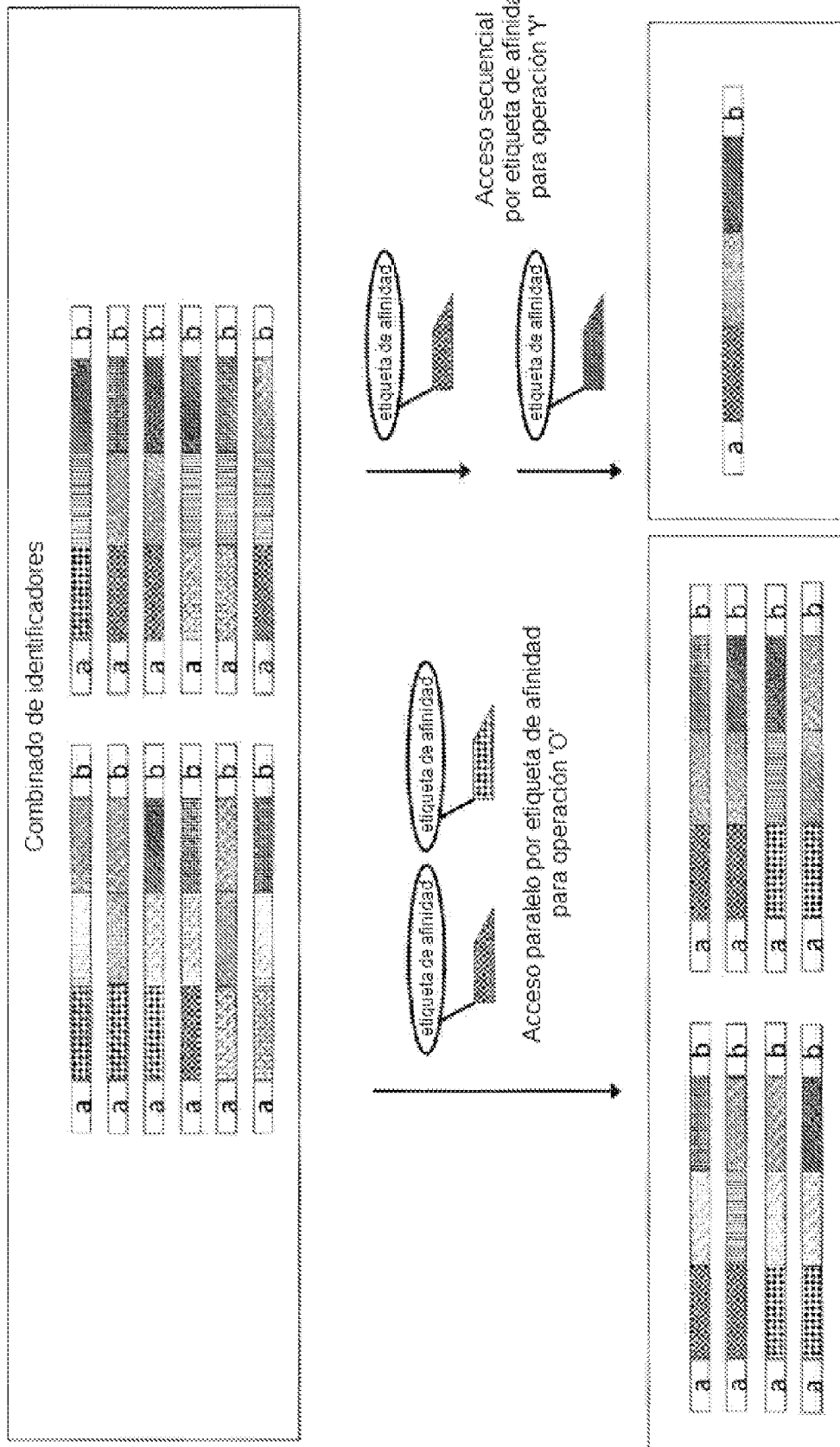


FIG. 17C

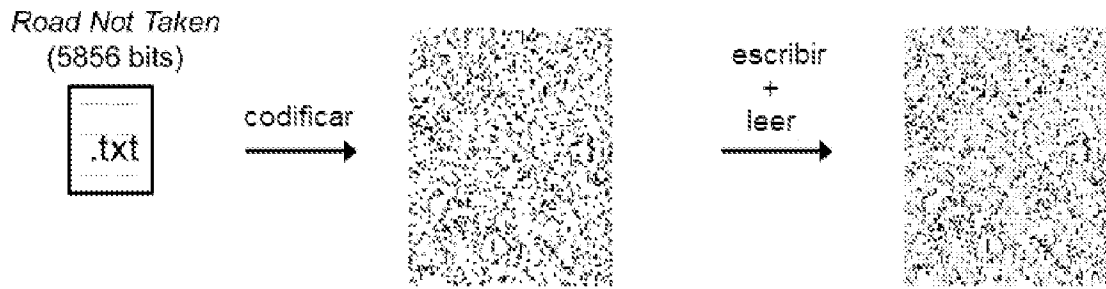


FIG. 18A

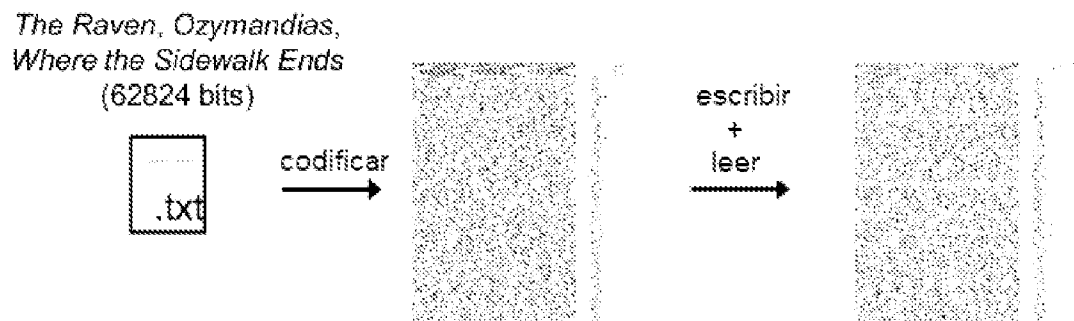


FIG. 18B

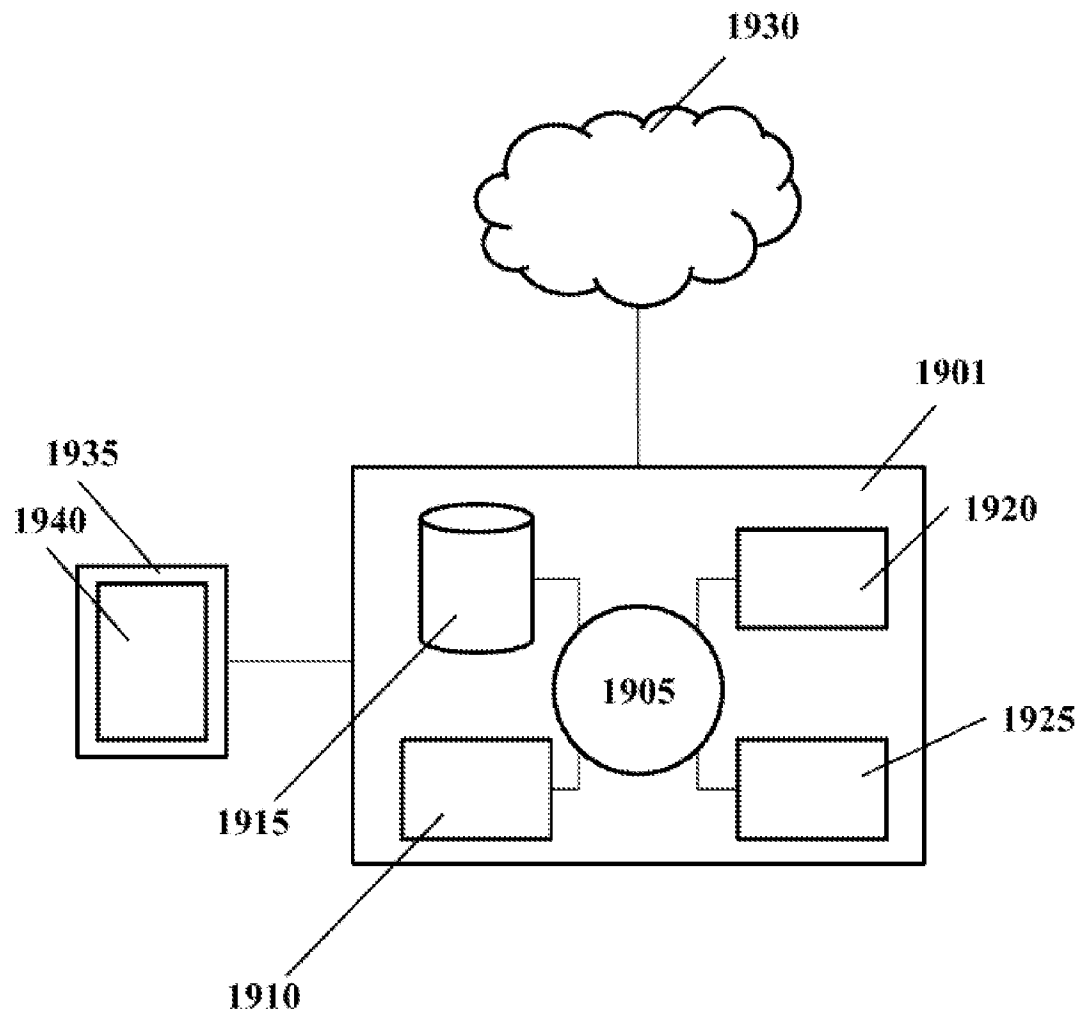


FIG. 19