

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 998 109**

51 Int. Cl.:

G16B 25/20 (2009.01)

G16B 25/00 (2009.01)

G16B 30/00 (2009.01)

C12Q 1/6811 (2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **10.08.2018** **PCT/KR2018/009178**

87 Fecha y número de publicación internacional: **14.02.2019** **WO19031916**

96 Fecha de presentación y número de la solicitud europea: **10.08.2018** **E 18844370 (9)**

97 Fecha y número de publicación de la concesión europea: **20.11.2024** **EP 3665688**

54 Título: **Métodos de preparación de oligonucleótidos para detectar secuencias diana de ácido nucleico con una cobertura máxima**

30 Prioridad:

11.08.2017 KR 20170102502

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

19.02.2025

73 Titular/es:

SEEGENE, INC. (100.00%)
8FL, 9FL Taewon Bldg., 91, Ogeum-ro, Songpa-gu
Seoul 05548, KR

72 Inventor/es:

JANG, MI HYUN

74 Agente/Representante:

ARIAS SANZ, Juan

Observaciones:

Véase nota informativa (Remarks, Remarques o Bemerkungen) en el folleto original publicado por la Oficina Europea de Patentes

ES 2 998 109 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

Métodos de preparación de oligonucleótidos para detectar secuencias diana de ácido nucleico con una cobertura máxima

Antecedentes de la invención

Campo de la invención

La presente invención se refiere a tecnologías de preparación de un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima, y para determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia.

Descripción de la técnica relacionada

Se ha desarrollado una variedad de técnicas para detectar moléculas diana de ácido nucleico de patógenos y para identificar estas moléculas diana de ácido nucleico, y estas se denominan colectivamente "diagnóstico molecular". La mayoría de las técnicas de diagnóstico molecular utilizan oligonucleótidos, tales como cebadores y sondas hibridables con moléculas diana de ácido nucleico.

Hasta el momento las tecnologías de diagnóstico molecular han realizado muchos avances. Sin embargo, todavía existen retos técnicos que no se han resuelto en el diagnóstico de patógenos que presentan genomas que muestran diversidad genética o variabilidad genética.

Se ha informado de diversidad genética o variabilidad genética en diversos genomas. En particular, se observa y ocurre diversidad genética más frecuentemente en los genomas víricos (Bastien N. et al., Journal of Clinical Microbiology, 42:3532(2004); Peret T.C. et al., Journal of Infectious Diseases, 185:1660(2002); Ebihara T. et al., Journal of Clinical Microbiology, 42:126(2004); Jenny-Avital E.R. et al., Clinical Infectious Diseases, 32:1227(2001); Duffy S. et al., Nat. Rev. Genet. 9(4):267-76(2008); Tong Y.G. et al., Nature 22:526(2015)).

En la detección de un patógeno con diversidad genética, el diseño de oligonucleótidos que considere una determinada secuencia de una molécula diana de ácido nucleico de dicho patógeno es muy probable que conduzca a falsos resultados negativos. De esta manera, a fin de determinar si está presente un determinado patógeno en una muestra problema, deberían diseñarse sondas o cebadores que consideren todas las secuencias de ácido nucleico o tantas secuencias de ácido nucleico como resulte posible, de diversidad genética conocida para una molécula diana de ácido nucleico de dicho patógeno determinado. Con el fin de detectar una molécula diana de ácido nucleico que muestre dicha diversidad genética, se han desarrollado principalmente dos enfoques.

El primer método detecta una molécula diana de ácido nucleico utilizando una pluralidad de oligonucleótidos que se hibridan con una pluralidad de secuencias de ácido nucleico de una molécula diana de ácido nucleico que muestra diversidad genética. Por ejemplo, en el caso de que se utilice como diana el gen M del virus de la influenza A, todas las secuencias de ácido nucleico conocidas del gen M están alineadas y se diseñan sondas que son capaces de cubrir la totalidad de estas secuencias de ácido nucleico. En este caso, debido a que una única sonda no puede cubrir todos los genes M de diversas secuencias, se diseña una pluralidad de sondas (sondas con posiciones de sondeo diferentes unas de otras).

El segundo método es diseñar un oligonucleótido degenerado. Normalmente, se encuentra una región en todas las secuencias de ácido nucleico de un determinado gen que presenta diversidad genética, y el gen determinado se detecta con la cobertura de interés utilizando un cebador o sonda degenerado (que incluye una base degenerada en un sitio de variación) que se hibrida con la región (ver las patentes US n.º 8.735.063, n.º 8.318.423 y n.º 8.685.649).

Para el segundo enfoque, es muy importante aplicar óptimamente bases degeneradas al oligonucleótido utilizado para detectar diversas secuencias de ácido nucleico de una molécula diana de ácido nucleico que presenta diversidad. Considerando la conveniencia, eficiencia y economía del análisis, resulta deseable detectar una molécula diana de ácido nucleico con una cobertura diana máxima utilizando el oligonucleótido en el que se aplican óptimamente las bases degeneradas.

Convencionalmente, con el fin de detectar diversas secuencias de ácido nucleico de una molécula diana de ácido nucleico con diversidad genética, los investigadores han determinado las posiciones de introducción y el número de bases degeneradas para la máxima cobertura de una pluralidad de secuencias de ácido nucleico mediante la aplicación de bases degeneradas a una sonda de una manera secuencial o aleatoria.

Kelvin Li et al., Virology Journal, vol. 9, n.º 1, página 261, dan a conocer un diseño automatizado de cebador de PCR degenerado para la secuenciación de alto rendimiento que mejora la eficiencia de la secuenciación vírica, la construcción de una secuencia de consenso se lleva a cabo con ANDES.

Kelvin et al., BMC Research Notes, vol. 3, n.º 1, página 199, describe el programa ANDES, que es una herramienta de tipo estadístico para comparar las variaciones dentro de una muestra o entre múltiples muestras. Se basa en la alineación de múltiples secuencias (AMS) y se obtiene un perfil de posiciones a partir de la misma. La AMS se lleva a cabo siempre mediante consideración de las secuencias en referencia unas a otras. Se calcula la desviación cuadrática media (RMSD, por sus siglas en inglés) para la visualización de las diferencias entre alineaciones. Se calculan las entropías de Shannon como una medida de la incertidumbre.

Puede proponerse el enfoque convencional, en el que se aplica secuencial o aleatoriamente un número limitado de bases degeneradas en determinadas posiciones de sondas para determinar la cobertura máxima de las secuencias, en el caso de que el número de secuencias diana de ácido nucleico sea pequeño. Sin embargo, en el caso de que el número de secuencias diana de ácido nucleico sea mucho mayor y se requieran oligonucleótidos degenerados que cubran al máximo las secuencias, el enfoque convencional adolece de graves desventajas a la luz del hecho de que no solo se requiere mucho tiempo sino que también presenta una exactitud reducida.

Según el mejor conocimiento del presente inventor, no existe ninguna técnica anterior para resolver el problema técnico que resuelve la presente invención mediante una lógica de optimización, es decir, (i) preparar un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima, y (ii) determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia.

Descripción resumida de la invención

El presente inventor ha llevado a cabo investigaciones intensivas para desarrollar tecnologías para aplicar de manera óptima bases degeneradas y/o bases universales a oligonucleótidos para cubrir al máximo una pluralidad de secuencias diana de ácido nucleico, entre otros, una pluralidad de secuencias de ácido nucleico (particularmente, secuencias diana de ácido nucleico) de una molécula diana de ácido nucleico que muestra diversidad genética, con una velocidad y exactitud mejoradas. Como resultado, el presente inventor ha desarrollado lógica de optimización para una introducción óptima de bases degeneradas y/o bases universales en un oligonucleótido (p. ej., una sonda y un cebador) utilizado para detectar una pluralidad de secuencias diana de ácido nucleico, en un enfoque completamente diferente de los métodos convencionales, es decir, métodos empíricos y manuales. Además, el presente inventor ha encontrado que la lógica de optimización puede utilizarse en (i) la preparación de un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presenta similitud de secuencia con una cobertura diana máxima, y (ii) la determinación de una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presenta similitud de secuencia.

De acuerdo con lo anteriormente expuesto, es un objetivo de la presente invención proporcionar un método para preparar un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima.

Es otro objetivo de la presente invención proporcionar un método para determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia.

Es todavía otro objetivo de la presente invención proporcionar un medio de almacenamiento legible por ordenador que contiene instrucciones para configurar un procesador para llevar a cabo un método de preparación de un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima.

Es otro objetivo de la presente invención proporcionar un medio de almacenamiento legible por ordenador que contiene instrucciones para configurar un procesador para llevar a cabo un método para determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia.

Otros objetivos y ventajas de la presente invención resultarán evidentes a partir de la descripción detallada siguiente, considerada junto con las reivindicaciones adjuntas y dibujos.

Breve descripción de los dibujos

La Fig. 1 muestra una matriz binaria ($A=\{a_{ij}, a_{ij} \in \{0,1\}\}$) que consiste en un valor coincidente y un valor no coincidente que indica si las sondas presentan o no diversos patrones de secuencia coinciden con una sonda de referencia (oligonucleótido de referencia) según una realización de la presente invención.

La Fig. 2 es un diagrama de flujo que representa un procedimiento de preparación de un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima según una realización del presente invención.

La Fig. 3 es un diagrama de flujo que muestra un procedimiento de determinación de una región de sondeo en una pluralidad de secuencias diana de ácido nucleico con similitud de secuencia según una realización de la presente invención.

Descripción detallada de la presente invención

El propósito técnico de la presente invención es realizar eficientemente decisiones a las que podrían enfrentarse los analistas en la detección simultánea de una pluralidad de secuencias diana de ácido nucleico, particularmente una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia, particularmente (i) determinar una introducción óptima de bases degeneradas y/o bases universales en un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico con una cobertura diana máxima, y (ii) determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico.

Los inventores han desarrollado dos lógicas de optimización ideales para dichas decisiones. Dicha lógica comparte características comunes y, por lo tanto, la presente invención se divide en gran medida en dos aspectos. A continuación en el presente documento se proporciona una descripción detallada de la presente invención.

I. Preparación de un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico con una cobertura diana máxima.

En un aspecto de la presente invención se proporciona un método según la reivindicación 1 para preparar un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima, que comprende: (a) seleccionar cada una de las secuencias con similitud de secuencia en la misma región de sondeo de una pluralidad de secuencias diana de ácido nucleico como secuencia de sondeo o una secuencia oligonucleotídica para preparar una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos, (b) seleccionar una de la pluralidad de secuencias de sondeo a modo de secuencia de sondeo de referencia o una de la pluralidad de oligonucleótidos a modo de oligonucleótido de referencia, y (c) introducir una base degenerada y/o base universal en un número permisible predeterminado en la secuencia de sondeo de referencia o el oligonucleótido de referencia en una o más posiciones no coincidentes entre (i) la secuencia de sondeo de referencia o el oligonucleótido de referencia y (ii) la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos que van a hacerse coincidir al máximo con la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos, y seleccionar como secuencia de un oligonucleótido, la secuencia de sondeo de referencia o una secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una de la base degenerada y/o la base universal, preparando de esta manera un oligonucleótido con una cobertura diana máxima.

Un primer aspecto de la presente invención se refiere a un método según la reivindicación 1 para preparar una introducción óptima de una base degenerada y/o base universal en un número permisible predeterminado en un oligonucleótido (p. ej., una sonda y un cebador) utilizado para detectar una pluralidad de secuencias diana de ácido nucleico con una cobertura diana máxima. En otras palabras, el primer método de la presente invención es un método de optimización que maximiza la cobertura de la diana de un oligonucleótido para la secuencia diana de ácido nucleico que va a detectarse, dado un número limitado de base degeneradas y/o bases universales que van a introducirse en el oligonucleótido.

La expresión “un número permisible predeterminado” o “un número limitado” utilizada en el presente documento en referencia al número de bases degeneradas y/o bases universales se refiere a la detección de una pluralidad de secuencias diana de ácido nucleico con una cobertura diana máxima, mediante la introducción de un número permisible predeterminado o un número inferior, o un número limitado o inferior de bases degeneradas y/o universales en el oligonucleótido. Por ejemplo, el número limitado de bases degeneradas y/o bases universales que va a introducirse en el oligonucleótido de 3 representa la detección de una pluralidad de secuencias diana de ácido nucleico con una cobertura diana máxima mediante la introducción de tres o menos bases degeneradas y/o bases universales en el oligonucleótido.

La Fig. 2 es un diagrama de flujo de los procedimientos para llevar a cabo la invención de acuerdo con una realización de la invención. El método de la presente invención se describe en referencia de la Fig. 2:

Etapas (a): *preparación de una pluralidad de secuencias de sondeo u oligonucleótidos* (110).

En primer lugar, se prepara una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos mediante la selección de cada una de las secuencias con similitud de secuencia en la misma región de sondeo de una pluralidad de secuencias diana de ácido nucleico como secuencia de sondeo o secuencia oligonucleotídica. Se selecciona una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos como una secuencia de sonda de referencia o un oligonucleótido de referencia en la etapa (b), y sus secuencias son factores considerables al determinar una introducción óptima de bases degeneradas y/o bases universales en la secuencia de sonda de referencia u oligonucleótido de referencia en la etapa (c).

La expresión utilizada en el presente documento “molécula diana de ácido nucleico”, “molécula diana” o “ácido nucleico diana” se refiere a una molécula de nucleótidos en un organismo que se pretende detectar. Generalmente, la molécula diana de ácido nucleico presenta un nombre determinado e incluye todo un genoma y todas las moléculas de nucleótidos que forman un genoma (p. ej., gen, pseudogén, molécula de secuencia no codificante, región no traducida y algunas regiones del genoma).

La molécula diana de ácido nucleico incluye, por ejemplo, un ácido nucleico de célula procariótica (p. ej., *Mycoplasma pneumoniae*, *Chlamydomonas pneumoniae*, *Legionella pneumophila*, *Haemophilus influenzae*, *Streptococcus pneumoniae*, *Bordetella pertussis*, *Bordetella parapertussis*, *Neisseria meningitidis*, *Listeria monocytogenes*, *Streptococcus agalactiae*, *Campylobacter*, *Clostridium difficile*, *Clostridium perfringens*, *Salmonella*, *Escherichia coli*, *Shigella*, *Vibrio*, *Yersinia enterocolitica*, *Aeromonas*, *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, *Trichomonas vaginalis*, *Mycoplasma hominis*, *Mycoplasma genitalium*, *Ureaplasma urealyticum*, *Ureaplasma parvum*, *Mycobacterium tuberculosis*), un ácido nucleico de célula eucariótica (p. ej., protozoo y animal parasitario, hongo, levadura, planta superior, animal inferior y animal superior, incluyendo mamífero y ser humano), ácido nucleico vírico o ácido nucleico de viroide. Entre los parásitos de las células eucarióticas se incluyen, por ejemplo, *Giardia lamblia*, *Entamoeba histolytica*, *Cryptosporidium*, *Blastocystis hominis*, *Dientamoeba fragilis*, *Cyclospora cayetanensis*. Entre los ejemplos de tales virus se incluye virus influenza A (gripe A), virus influenza B (gripe B), virus A sincitial respiratorio (VSR A), virus B sincitial respiratorio (VSR B), virus parainfluenza 1 (VPI 1), virus parainfluenza 2 (VPI 2), virus parainfluenza 3 (VPI 3), virus parainfluenza 4 (VPI 4), metapneumovirus (MPV), enterovirus humano (EVH), bocavirus humano (BoVH), rinovirus humano (RVH), coronavirus y adenovirus, que causan enfermedades respiratorias; norovirus, rotavirus, adenovirus, astrovirus y sapovirus, que causan trastornos gastrointestinales. El virus incluye, además, por ejemplo, papilomavirus humano (PVH), coronavirus relacionado con el síndrome respiratorio de Oriente Medio (CoV-SROM), virus dengue, virus del herpes simple (VHS), virus del herpes humano (VHH), virus de Epstein-Barr (VEB), virus de la varicela zóster (VVZ), citomegalovirus (CMV), VIH, virus de la hepatitis y poliovirus.

Las expresiones utilizadas en el presente documento “secuencia diana de ácido nucleico” o “secuencia diana” representa una molécula diana de ácido nucleico como una determinada secuencia.

Una molécula diana de ácido nucleico, p. ej., un gen diana, puede presentar una determinada secuencia diana de ácido nucleico; en caso contrario, para una molécula diana de ácido nucleico que muestra diversidad genética o variabilidad genética, puede presentar una pluralidad de secuencias diana de ácido nucleico con diversidad. En el caso de que se pretenda detectar una molécula diana de ácido nucleico que muestra diversidad genética sin ningún resultado falso negativo, normalmente se requiere la combinación de una pluralidad de oligonucleótidos o un oligonucleótido en el que se han introducido bases degeneradas y/o bases universales que es capaz de cubrir una pluralidad de secuencias diana de ácido nucleico que presentan diversidad.

La pluralidad de secuencias diana de ácido nucleico en la presente invención son secuencias diana de ácido nucleico que presentan similitud de secuencia. En particular, las secuencias diana de ácido nucleico que presentan similitud de secuencia pueden ser una pluralidad de secuencias diana de ácido nucleico de una molécula diana de ácido nucleico o una pluralidad de secuencias diana de ácido nucleico de dos o más moléculas diana de ácido nucleico.

Según una realización, la pluralidad de secuencias diana de ácido nucleico en la presente invención son una pluralidad de secuencias de ácido nucleico que presentan similitud de secuencia para una molécula diana de ácido nucleico que presenta diversidad genética.

Por ejemplo, la pluralidad de secuencias diana de ácido nucleico utilizadas en la presente invención son una pluralidad de secuencias de ácido nucleico que presentan similitud de secuencia para una molécula diana de ácido nucleico que muestra diversidad genética, tal como una secuencia de genoma vírico. Por ejemplo, en el caso de que se pretenda detectar el virus influenza A y se determine el gen M como una molécula diana de ácido nucleico, pueden utilizarse secuencias diana de ácido nucleico con diversidad del gen M del virus influenza A. El virus influenza A incluye una variedad de subtipos y variantes, y sus secuencias genómicas son diferentes entre sí. Por lo tanto, en el caso de que se pretenda detectar el virus influenza A sin ningún resultado falso negativo, debe diseñarse un oligonucleótido que considere diversas secuencias diana de ácido nucleico de una molécula diana de ácido nucleico del virus influenza A originada a partir de dicha diversidad genética.

Más particularmente, la pluralidad de secuencias diana de ácido nucleico son una secuencia genómica completa, una secuencia parcial de un genoma, o una pluralidad de secuencias de ácido nucleico de un gen de virus o bacteria que presenta diversidad genética.

Según una realización de la presente invención, la pluralidad de secuencias diana de ácido nucleico son una pluralidad de secuencias de ácido nucleico correspondientes a homólogos de organismos que presentan la misma función, la misma estructura, o el mismo nombre del gen. El organismo se refiere a un organismo perteneciente a un género, especie, subespecie, subtipo, genotipo, serotipo, cepa, aislado o cultivar. Entre los homólogos se incluyen moléculas de proteínas y ácidos nucleicos. En la presente realización, se utiliza una pluralidad de secuencias de ácido nucleico de moléculas biológicas homólogas (p. ej., proteínas o ácidos nucleicos) de una pluralidad de organismos que presentan la misma función (p. ej., una función biológica de una proteína codificada por una secuencia de ácido

nucleico), la misma estructura (p. ej., una estructura terciaria de una proteína codificada por una secuencia de ácido nucleico) o el mismo nombre de gen. Por ejemplo, puede considerarse una pluralidad de secuencias de ácido nucleico conocidas para el gen E5 del VPH tipo 16 como secuencia de ácido nucleico de aislados del VPH tipo 16. En el caso de que se utilice el gen E5 como molécula diana de ácido nucleico para detectar el VPH tipo 16, deben introducirse bases degeneradas y/o bases universales en el oligonucleótido para poder cubrir una pluralidad de secuencias de ácido nucleico con diversidad del gen E5 del VPH tipo 16.

Según una realización, la secuencia diana de ácido nucleico comprende secuencias de ácido nucleico pertenecientes a una subclase de cualquier clasificación biológica (p. ej., género, especie, subtipo, genotipo, serotipo y subespecie). Por ejemplo, en el caso de que la secuencia diana de ácido nucleico sea VPH tipo 16, la secuencia diana de ácido nucleico puede comprender secuencias de ácido nucleico pertenecientes a esa subclase.

Según una realización de la presente invención, la pluralidad de secuencias diana de ácido nucleico son por lo menos 3, por lo menos 5, por lo menos 10, por lo menos 20, por lo menos 30, por lo menos 40, por lo menos 50, por lo menos 60, por lo menos 70, por lo menos 80, por lo menos 90 o por lo menos 100 secuencias de ácido nucleico.

Puede proporcionarse una pluralidad de secuencias diana de ácido nucleico utilizando diversas bases de datos de secuencias. Por ejemplo, puede recogerse una pluralidad de secuencias diana de ácido nucleico deseadas y proporcionarse a partir de una base de datos públicamente accesible, tal como GenBank, la base de datos de secuencias del European Molecular Biology Laboratory (EMBL), y el DNA DataBank of Japan (DDBJ).

Según una realización, se alinea una pluralidad de secuencias diana de ácido nucleico previamente a la etapa (a). La alineación de secuencias diana de ácido nucleico puede llevarse a cabo de acuerdo con diversos métodos (p. ej., alineación global y alineación local) y algoritmos conocidos de la técnica.

Se describen diversos métodos y algoritmos de alineación en Smith y Waterman, Adv. Appl. Math. 2:482(1981); Needleman y Wunsch, J. Mol. Biol. 48:443(1970); Pearson y Lipman, Methods in Mol. Biol. 24:307-31(1988); Higgins y Sharp, Gene 73:237-44(1988); Higgins y Sharp, CABIOS 5:151-3(1989); Corpet et al., Nuc. Acids Res. 16:10881-90 (1988); Huang et al., Comp. Appl. BioSci. 8:155-65(1992) y Pearson et al., Meth. Mol. Biol. 24:307-31(1994). La herramienta de búsqueda de alineaciones locales básicas (BLAST, por sus siglas en inglés) del NCBI (Altschul et al., J. Mol. Biol. 215:403-10(1990)) es accesible del NCBI (National Center for Biological Information) y puede utilizarse junto con programas de análisis de secuencias, tales como blastn, blastm, blastx, tblastn y tblastx en internet. BLSAT está disponible en <http://www.ncbi.nlm.nih.gov/BLAST/>. Puede encontrarse una comparación de similitudes de secuencia utilizando dicho programa en http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html.

Según la presente invención, se alinea la pluralidad de secuencias diana de ácido nucleico y se selecciona cada una de las secuencias que presenta similitud de secuencia en la misma región de sondeo, a modo de secuencia de sondeo o una secuencia oligonucleotídica.

La expresión utilizada en el presente documento "región de sondeo" se refiere a una región de la secuencia diana de ácido nucleico que resulta adecuada para la hibridación específica de diana de oligonucleótidos, tales como cebadores y sondas, y la expresión comprende una región de cebado.

Por ejemplo, en el caso de que haya 10 secuencias diana de ácido nucleico entre A y J, las secuencias diana de ácido nucleico entre A y J están alineadas y se seleccionan secuencias de la misma región de una longitud determinada que presenten similitud de secuencia a modo de secuencias de sondeo de las secuencias diana de ácido nucleico entre A y J.

Según una realización de la invención, las regiones de sondeo con similitud de secuencia adecuadas para la selección como las secuencias de sondeo presentan una similitud de secuencia de 20 % a 100 %, 30 % a 100 %, 40 % a 100 %, 50 % a 100 %, 60 % a 100 %, 70 % a 100 %, 80 % a 100 % o 90 % a 100 %.

Cada una de las secuencias que presenta similitud de secuencia en la misma región de sondeo puede seleccionarse como una secuencia de sondeo para cada una de una pluralidad de secuencias diana de ácido nucleico y también puede seleccionarse como una secuencia oligonucleotídica. En la presente especificación, los oligonucleótidos, excepto el oligonucleótido que muestra una cobertura diana máxima preparado finalmente en la etapa (c), son aquellos utilizados para la preparación del oligonucleótido que muestra la cobertura diana máxima. Las secuencias de dichos oligonucleótidos pueden comprender secuencias que se hibridan con secuencias de sondeo o que se hibridan con sus secuencias complementarias. En el caso de que los oligonucleótidos sean secuencias que se hibridan con secuencias complementarias a secuencias de sondeo, las secuencias de sondeo pueden ser secuencias oligonucleotídicas. Es decir, el oligonucleótido que presenta la máxima cobertura diana finalmente preparado puede proporcionarse utilizando una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos.

Según una realización de la presente invención, cada una de la pluralidad de secuencias de sondeo u oligonucleótidos presenta un índice único.

Etapas (b): selección de una secuencia de sondeo de referencia u oligonucleótido (120).

A continuación, se selecciona una de la pluralidad de secuencias de sondeo o uno de la pluralidad de oligonucleótidos a modo de secuencia de sondeo de referencia u oligonucleótido de referencia, respectivamente.

La secuencia de sondeo de referencia seleccionada u oligonucleótido de referencia es una referencia para determinar si la coincidencia debe ser con (i) la pluralidad de secuencias de sondeo o (ii) la pluralidad de oligonucleótidos de la etapa (c), y es el sujeto de introducción al introducir una base degenerada y/o base universal en una o más posiciones no coincidentes entre las secuencias.

Para este propósito, puede utilizarse una secuencia de sondeo de referencia o un oligonucleótido de referencia mediante la selección de una cualquiera de la pluralidad de secuencias de sondeo o una cualquiera de la pluralidad de oligonucleótidos.

Según una realización, la secuencia de sondeo de referencia es una secuencia de sondeo que presenta el mayor número de secuencias idénticas de la pluralidad de secuencias de sondeo, y el oligonucleótido de referencia es un oligonucleótido que presenta el número más grande de secuencias idénticas de la pluralidad de oligonucleótidos.

Debido a que el oligonucleótido proporcionado por la presente invención se prepara mediante la introducción óptima de una base degenerada y/o base universal para detectar una pluralidad de secuencias diana de ácido nucleico con la máxima cobertura diana, resulta deseable que la secuencia de sondeo con el número más grande de secuencias idénticas de la pluralidad de secuencias de sondeo o el oligonucleótido que presenta el número más grande de secuencias idénticas de la pluralidad de oligonucleótidos se seleccione como la secuencia de sondeo de referencia o el oligonucleótido de referencia.

Etapas (c): preparación de un oligonucleótido con la máxima cobertura diana (130).

A continuación, se prepara un oligonucleótido con una cobertura diana máxima mediante la introducción de una base degenerada y/o base universal en un número permisible predeterminado en la secuencia de sondeo de referencia o el oligonucleótido de referencia en una o más posiciones no coincidentes entre (i) la secuencia de sondeo de referencia o el oligonucleótido de referencia, y (ii) la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos que coincidirán al máximo con la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos, y seleccionar como secuencia de un oligonucleótido, la secuencia de sondeo de referencia o la secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una de base degenerada y/o base universal.

En la presente invención, para determinar la posición o posiciones en las que se ha introducido por lo menos una de base degenerada y/o base universal, la referencia para la comparación es la secuencia de sondeo de referencia o el oligonucleótido de referencia y el sujeto de comparación es (ii) una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos. De esta manera, en la presente invención, la comparación de una coincidencia o no coincidencia para introducir la base degenerada y/o base universal puede llevarse a cabo de cuatro maneras, particularmente, entre una secuencia de sondeo de referencia y una pluralidad de secuencias de sondeo, entre una secuencia de sondeo de referencia y una pluralidad de oligonucleótidos, entre un oligonucleótido de referencia y una pluralidad de secuencias de sondeo, o entre un oligonucleótido de referencia y una pluralidad de oligonucleótidos.

La expresión "no coincidencia" utilizada en el presente documento en referencia a la posición o posiciones de introducción de una base degenerada y/o base universal se refiere a que dos bases correspondientes a la misma posición de dos secuencias no son idénticas en el caso de que la dirección de las dos secuencias que van a compararse no sean iguales, y a que las dos bases correspondientes de las dos secuencias no son complementarias en el caso de que la dirección sea diferente. Por ejemplo, en el caso de que una secuencia del oligonucleótido de referencia como referencia para la comparación de no coincidencias y la pluralidad de secuencias de sondeo como sujeto de comparación presenten ambas una dirección de 5' a 3', y en la misma posición el oligonucleótido de referencia presente "A" y la secuencia de sondeo como sujeto de la comparación presente una base diferente de "A", la posición es una posición de no coincidencia. Alternativamente, en el caso de que una secuencia del oligonucleótido de referencia presente una dirección de 5' a 3' y la pluralidad de secuencias de sondeo como sujeto de comparación presente una dirección de 3' a 5', y en la misma posición el oligonucleótido de referencia presente "A" y la secuencia de sondeo presente una base diferente de "T", la posición es una posición de no coincidencia.

En la presente invención, la base degenerada y/o base universal en un número permisible predeterminado se introduce en la secuencia de sondeo de referencia o el oligonucleótido de referencia en la posición o posiciones no coincidentes entre (i) la secuencia de sondeo de referencia o el oligonucleótido de referencia y (ii) la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos. El sujeto de introducción de la base degenerada y/o base universal en la presente invención es la secuencia de sondeo de referencia o el oligonucleótido de referencia.

El número permisible predeterminado de bases degeneradas y/o bases universales que se introduce en la presente invención es particularmente 7 o menos, 5 o menos, 4 o menos, o 3 o menos. La tasa de utilización de las bases degeneradas y/o bases universales que se introducen en la presente invención es particularmente 25 % o inferior, 20

% o inferior, 18 % o inferior, 16 % o inferior, 14 % o inferior, 12 % o inferior, 10 % o inferior, 8 % o inferior, o 6 % o inferior. La tasa de utilización de las bases degeneradas y/o bases universales indica la proporción de una base degenerada y/o base universal en el total de nucleótidos del oligonucleótido en el que se introduce la base degenerada y/o universal.

5 En la presente invención, la base degenerada y/o base universal se introduce en la secuencia de sondeo de referencia o en el oligonucleótido de referencia para que coincidan al máximo con la pluralidad de secuencias de sondeo o con la pluralidad de oligonucleótidos.

10 La expresión "coincidencia máxima" utilizada en el presente documento en referencia a la introducción de la base degenerada y/o base universal significa que se maximiza el número de la pluralidad de secuencias de sondeo o de la pluralidad de oligonucleótidos que presentan la misma secuencia que la secuencia de sondeo de referencia o que el oligonucleótido de referencia en la que se introduce la base degenerada y/o base universal, y puede expresarse la máxima coincidencia como una proporción o porcentaje del número total de la pluralidad de secuencias de sondeo o de la pluralidad de oligonucleótidos.

15 En la presente invención, en el caso de que la secuencia de sondeo de referencia o el oligonucleótido de referencia coincida al máximo con la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos por la introducción de bases degeneradas y/o bases universales en un número permisible predeterminado, la secuencia de sondeo de referencia o una secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una base degenerada y/o base universal se selecciona como una secuencia de un oligonucleótido con la máxima cobertura diana.

20 La secuencia de oligonucleótido seleccionada de la secuencia de sondeo de referencia en la que se introduce la base degenerada y/o base universal puede utilizarse para detectar una pluralidad de secuencias diana de ácido nucleico correspondiente a una pluralidad de secuencias de sondeo que coinciden al máximo con la secuencia de sondeo de referencia en la que se ha introducido la base degenerada y/o base universal. La secuencia oligonucleotídica seleccionada de la secuencia oligonucleotídica de referencia en la que se ha introducido la base degenerada y/o base universal puede utilizarse para detectar una pluralidad de secuencias diana de ácido nucleico correspondiente a una pluralidad de secuencias del oligonucleótido que coinciden al máximo con el oligonucleótido de referencia en el que se ha introducido la base degenerada y/o base universal.

25 La selección como la secuencia del oligonucleótido, la secuencia de sondeo de referencia o la secuencia del oligonucleótido de referencia en la que se introduce por lo menos una base degenerada y/o base universal puede comprender tanto (i) la selección como la secuencia del oligonucleótido, de la secuencia de sondeo de referencia o la secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una base degenerada y/o base universal, y (ii) la selección como la secuencia del oligonucleótido, de una secuencia complementaria a la secuencia de sondeo de referencia o una secuencia complementaria a la secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una base degenerada y/o base universal.

30 La expresión "cobertura de la diana" utilizada en el presente documento en referencia a un oligonucleótido preparado mediante la presente invención se refiere a un valor que indica una proporción de la pluralidad de secuencias diana de ácido nucleico con la que se hibrida específicamente el oligonucleótido preparado.

35 Particularmente, la expresión "cobertura de la diana" se refiere a la proporción de la secuencia diana de ácido nucleico respecto a la pluralidad de secuencias diana de ácido nucleico con la que se hibrida el oligonucleótido preparado con una complementariedad sustancial (particularmente, con perfecta complementariedad o con perfecta coincidencia). La cobertura de la diana puede expresarse como proporción o en porcentaje.

40 Según una realización de la presente invención, el oligonucleótido es una sonda y/o un cebador. El término "sonda" utilizado en el presente documento se refiere a una molécula de ácido nucleico de una sola cadena que comprende una parte o partes que son complementarias a una secuencia diana de ácido nucleico. El término "cebador" tal como se utiliza en el presente documento se refiere a un oligonucleótido, que es capaz de actuar como un punto de inicio de síntesis al someterlo a condiciones en las que se induce la síntesis del producto de extensión del cebador que es complementario a una secuencia diana de ácido nucleico, es decir, en la presencia de nucleótidos y un agente para la polimerización, tal como ADN polimerasa, y a una temperatura y pH adecuados.

45 Los oligonucleótidos preparados en la presente invención pueden presentar una estructura convencional de cebador y sonda que consiste en secuencias que se hibridan con una secuencia diana de ácido nucleico. Alternativamente, el oligonucleótido preparado en la presente invención puede presentar una estructura única. Por ejemplo, los oligonucleótidos preparados en la presente invención pueden presentar una estructura de cebador Scorpion, sonda de baliza molecular, cebador Sunrise, sonda HyBeacon, sonda de etiquetado, cebador o sonda DPO (documento n.º WO 2006/095981) y sonda PTO (documento n.º WO 2012/096523).

60 Los oligonucleótidos preparados en la presente invención pueden ser un oligonucleótido modificado, tal como un oligonucleótido que contiene una base degenerada y/o un oligonucleótido que contiene una base universal, en donde

las bases degeneradas y/o bases universales se introducen en un cebador o sonda convencional. Las expresiones "cebador convencional", "sonda convencional" y "oligonucleótido convencional" utilizadas en el presente documento se refiere a un cebador, sonda y oligonucleótido habitual en el que no se introduce una base degenerada o base no natural. Según una realización, en el caso de que se prepare en la presente invención un oligonucleótido que contiene una base degenerada o un oligonucleótido que contiene una base universal, por lo menos 50 %, por lo menos 60 %, por lo menos 70 %, por lo menos 80 %, por lo menos 90 % o por lo menos 95 % del oligonucleótido es oligonucleótido no modificado. La base degenerada incluye las diversas bases degeneradas conocidas de la técnica, tales como las siguientes: R: A o G; Y: C o T; S: G o C; W: A o T; K: G o T; M: A o C; B: C o G o T; D: A o G o T; H: A o C o T; V: A o C o G; N: A o C o G o T. La base universal incluye las diversas bases universales siguientes conocidas de la técnica: desoxiinosina, inosina, 7-deaza-2'-desoxiinosina, 2-aza-2'-desoxiinosina, 2'-OMe-inosina, 2'-F-inosina, desoxi-3-nitropirrol, 3-nitropirrol, 2'-OMe-3-nitropirrol, 2'-F-3-nitropirrol, 1-(2'-desoxi-beta-D-ribofuranosil)-3-nitropirrol, desoxi-5-nitropirrol, 5-nitroindol, 2'-OMe-5-nitroindol, 2'-F-5-nitroindol, desoxi-4-nitrobencimidazol, 4-nitrobencimidazol, desoxi-4-aminobencimidazol, 4-aminobencimidazol, desoxinebularina, 2'-F-nebularina, 2'-F-4-nitrobencimidazol, PNA-5-introindol, PNA-nebularina, PNA-inosina, PNA-4-nitrobencimidazol, PNA-3-nitropirrol, morfolino-5-nitroindol, morfolino-nebularina, morfolino-inosina, morfolino-4-nitrobencimidazol, morfolino-3-nitropirrol, fosforamidato-5-nitroindol, fosforamidato-nebularina, fosforamidato-inosina, fosforamidato-4-nitrobencimidazol, fosforamidato-3-nitropirrol, 2'-O-metoxietil-inosina, 2'-O-metoxietil-nebularina, 2'-O-metoxietil-5-nitroindol, 2'-O-metoxietil-4-nitro-bencimidazol, 2'-O-metoxietil-3-nitropirrol, y combinaciones de los mismos. Más particularmente, la base universal es desoxiinosina, inosina o combinaciones de las mismas.

Según una realización, la base introducida para una cobertura diana máxima es una base degenerada. Entre los oligonucleótidos degenerados se incluyen una pluralidad de oligonucleótidos representados por oligonucleótidos degenerados. A menos que se indique específicamente en el presente documento, "oligonucleótidos degenerados" se refiere a un subgrupo que comprende una pluralidad de oligonucleótidos representados por oligonucleótidos degenerados, en lugar de un solo oligonucleótido.

Los oligonucleótidos preparados en la presente invención pueden ser sondas o cebadores que satisfacen por lo menos uno de los criterios adicionales siguientes: (i) un valor de T_f de entre 35 °C y 85 °C, (ii) una longitud de entre 15 y 50 nucleótidos, (iii) una secuencia de nucleótidos con un contenido de GC de entre 30 % y 80 %, (iv) el valor de ΔG en el que el oligonucleótido forma una estructura de horquilla es -8,0 kcal/mol o superior, (v) en el caso de que el oligonucleótido forme un homodímero, la proporción de nucleótidos consecutivos que participan en la formación del homodímero es de 65 % o inferior, (vi) en el caso de que el oligonucleótido forme un heterodímero con otro oligonucleótido, la proporción de nucleótidos que participan en la formación del heterodímero es de 70 % o inferior, y (vii) en el caso de que el oligonucleótido forme un heterodímero con otro oligonucleótido, la proporción de nucleótidos consecutivos que participan en la formación del heterodímero es de 65 % o inferior.

Según una realización de la presente invención, el método comprende, además, entre las etapas (a) y (b), (a-1) la agrupación de la pluralidad de secuencias de sondeo o de la pluralidad de oligonucleótidos según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, en donde la etapa (c) se lleva a cabo mediante la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la secuencia de sondeo de referencia o el oligonucleótido de referencia en la posición o posiciones no coincidentes entre la secuencia de sondeo de referencia o el oligonucleótido de referencia y la pluralidad de patrones de secuencia de máxima coincidencia con la pluralidad de patrones de secuencia, preparando de esta manera el oligonucleótido con la máxima cobertura diana.

Mediante la agrupación de la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, se vuelve posible reducir el número de comparaciones de las secuencias de sondeo u oligonucleótidos para comprobar si coinciden o no con la secuencia de sondeo de referencia o el oligonucleótido de referencia, reduciendo de esta manera el tiempo para proporcionar el oligonucleótido que presenta una cobertura diana máxima.

La etapa (c) anteriormente mencionada puede aplicarse a una pluralidad de patrones de secuencia de una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos.

Según una realización de la presente invención, cada uno de la pluralidad de patrones de secuencia presenta un índice único.

Según una realización, la secuencia de sondeo de referencia es una secuencia de sondeo de un patrón de secuencia que presenta el número más grande de secuencias de sondeo agrupadas en un patrón de secuencia de entre la pluralidad de patrones de secuencia, y el oligonucleótido de referencia es un oligonucleótido de un patrón de secuencia que presenta el número más grande de oligonucleótidos agrupados en un patrón de secuencia de la pluralidad de patrones de secuencia.

Una de las características de la presente invención es la aplicación de programación lineal como lógica de optimización, en la introducción óptima de una base degenerada y/o base universal en un número permisible

predeterminado en un oligonucleótido utilizado para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima.

La "programación lineal" utilizada en la presente invención es una técnica de optimización que maximiza o minimiza una función objetiva lineal, a la vez que satisface una fórmula de restricción lineal dada (incluyendo la fórmula de restricción de número no negativo de una variable, una variable ≥ 0). La programación lineal utiliza modelos matemáticos para expresar problemas, y tanto las fórmulas objetivas como las de restricción son formas lineales (es decir, funciones lineales).

Según la presente invención, la etapa (c) se lleva a cabo para conseguir la fórmula objetiva 1 siguiente, además de satisfacer las fórmulas de restricción 1 y 2 siguientes:

Fórmula objetiva 1:

$$\text{Máx.: } \sum_{i=1}^r x_i$$

en donde Máx.: representa maximización; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) de si selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos, e " r " es un índice de secuencias de sondeo u oligonucleótidos comprendido entre 1 y r ,

Fórmula de restricción 1:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en donde d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si se introduce la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; j es un índice de posiciones de una secuencia de sondeo de referencia o un oligonucleótido de referencia comprendido entre 1 y c ; y $D_{lim.}$ es un número limitado de bases degeneradas y/o bases universales introducidas en la secuencia de sondeo de referencia u oligonucleótido de referencia;

Fórmula de restricción 2:

$$x_i \leq \min\{a_{ij} + d_j\} \text{ para todo } i, j$$

en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos; a_{ij} es una constante binaria que representa si la i -ésima secuencia de sondeo es coincidente o no coincidente en la posición j -ésima con la secuencia de sondeo de referencia o el i -ésimo oligonucleótido es coincidente o no coincidente en la j -ésima posición con el oligonucleótido de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; $a_{ij} + d_j$ es una variable binaria que representa si la secuencia de sondeo de referencia con o sin la base degenerada y/o la base universal introducida es coincidente o no coincidente en la j -ésima posición con la i -ésima secuencia de sondeo o el oligonucleótido de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con el i -ésimo oligonucleótido; $\{a_{ij} + d_j\}$ representa un conjunto que incluye $a_{ij} + d_j$ como elementos; $\min\{a_{ij} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{ij} + d_j\}$ y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo u oligonucleótidos.

El problema que se va a resolver mediante la programación lineal en la presente invención es que debe introducirse un número limitado de bases degeneradas y/o bases universales en un oligonucleótido para representar una cobertura diana máxima.

Con el fin de determinar la máxima cobertura diana, es necesario seleccionar cada una de la pluralidad de secuencias de sondeo o de la pluralidad de oligonucleótidos para la máxima coincidencia con la secuencia de sondeo de referencia o el oligonucleótido de referencia.

De esta manera, en la presente invención, se adopta x_i como una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos. En la presente invención, i es un índice de secuencias de sondeo u oligonucleótidos comprendido entre 1 y r , la r es el índice de la última secuencia de sondeo u oligonucleótido.

Debido a que x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos, en el caso de que no se seleccione la i -ésima secuencia de sondeo o el i -ésimo oligonucleótido, se indica como el valor de no selección ($x_{no\ sel.}$), y en el caso de que se seleccione la i -ésima secuencia de sondeo o el i -ésimo oligonucleótido, se indica como el valor de selección ($x_{sel.}$). Particularmente, el valor de no selección ($x_{no\ sel.}$) es 0 y el valor de selección ($x_{sel.}$) es un valor diferente de 0. Más particularmente, el valor de no selección ($x_{no\ sel.}$) y el valor de selección ($x_{sel.}$) de x_i son 0 y 1, respectivamente.

Debido a que se ha de introducir un número limitado de bases degeneradas y/o bases universales en un oligonucleótido, es necesario determinar si introducir o no la base degenerada y/o base universal en qué posición o posiciones de la secuencia de sondeo de referencia o del oligonucleótido de referencia. De esta manera, se adopta d_j como una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si se introduce o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia. En la presente invención, j es un índice de posiciones de una secuencia de sondeo de referencia o un oligonucleótido de referencia comprendido entre 1 y 2, y c es el índice de la última posición en la secuencia de sondeo de referencia o del oligonucleótido de referencia.

El valor "c" no se encuentra particularmente limitado y puede ser, por ejemplo, un número entero entre 10 y 100, entre 10 y 80, entre 10 y 50, entre 10 y 40, entre 10 y 30, entre 15 y 100, entre 15 y 80, entre 15 y 50, entre 15 y 40, entre 15 y 30, entre 20 y 100, entre 20 y 80, entre 20 y 50, entre 20 y 40, entre 20 y 30, entre 25 y 100, entre 25 y 80, entre 25 y 50, entre 25 y 40, entre 25 y 30, entre 30 y 100, entre 30 y 80, entre 30 y 50, entre 30 y 40, entre 35 y 100, entre 35 y 80, entre 35 y 50 o entre 35 y 40.

Debido a que d_j es una variable binaria que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o del oligonucleótido de referencia, en el caso de que la base degenerada y/o la base universal no se introduzca en la j -ésima posición de la secuencia de sondeo de referencia o del oligonucleótido de referencia, se indica como el valor de no introducción ($d_{no\ int.}$), y en el caso de que se introduzca la base degenerada y/o la base universal en la j -ésima posición de la secuencia de sondeo de referencia o del oligonucleótido de referencia, se indica como el valor de introducción ($d_{int.}$). En particular, el valor de no introducción ($d_{no\ int.}$) es 0 y el valor de introducción ($d_{int.}$) es un valor diferente de 0. Más particularmente, el valor de no introducción ($d_{no\ int.}$) y el valor de introducción ($d_{int.}$) de d_j son 0 y 1, respectivamente.

Debido a que el oligonucleótido preparado en la presente invención debe presentar la máxima cobertura diana, el número de la pluralidad de secuencias de sondeo o de la pluralidad de oligonucleótidos coincidentes con la secuencia de sondeo de referencia o con el oligonucleótido de referencia en el que se ha introducido la base degenerada y/o la base universal debe ser el más alto posible. Por lo tanto, la suma del valor de selección y el valor de no selección para la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos se convierte en una función objetiva, y la función objetiva debe maximizarse. Por lo tanto, la fórmula objetiva en la presente invención puede expresarse como la fórmula objetiva 1.

Además, en la presente invención, mediante la introducción de un número limitado de bases degeneradas y/o bases universales en la secuencia de sondeo de referencia o en el oligonucleótido de referencia, la secuencia de sondeo de referencia o el oligonucleótido de referencia es necesario que presente la máxima coincidencia con una pluralidad de secuencias de sondeo o con una pluralidad de oligonucleótidos. Es decir, debe utilizarse un número limitado de bases degeneradas y/o bases universales para la máxima coincidencia o la máxima cobertura diana. Por lo tanto, en la presente invención, la fórmula de restricción 1 se establece de acuerdo con dicha condición de restricción.

En la fórmula de restricción 1, D_{lim} es un número limitado de bases degeneradas y/o bases universales introducidas en la secuencia de sondeo de referencia o en el oligonucleótido de referencia.

El número limitado de bases degeneradas y/o bases universales introducidas en la presente invención es particularmente 7, 5, 4 o 3.

En la lógica de optimización de la presente invención, la base degenerada y/o la base universal se introduce en la secuencia de sondeo de referencia o el oligonucleótido de referencia para que coincida al máximo con la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos. Por lo tanto, se requiere una fórmula de restricción relacionada con la coincidencia tras la introducción de la base degenerada y/o la base universal en la secuencia de sondeo de referencia o el oligonucleótido de referencia y con la selección de la i -ésima secuencia de sondeo de entre la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de entre la pluralidad de oligonucleótidos. Por este motivo, se establece la fórmula de restricción 2.

El valor de coincidencia o no coincidencia que representa la coincidencia o no coincidencia entre la secuencia de sondeo de referencia o el oligonucleótido de referencia en el que no se introduce la base degenerada y/o la base universal y una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos no es un variable, sino una constante. Por este motivo, se adopta a_{ij} , que es una constante binaria que representa si la i -ésima secuencia de sondeo es coincidente o no coincidente en la j -ésima posición respecto a la secuencia de sondeo de referencia o si el i -ésimo oligonucleótido es coincidente o no coincidente en la j -ésima posición respecto al oligonucleótido de referencia.

Particularmente, en el caso de que la j -ésima posición de la secuencia de sondeo de referencia antes de la introducción de la base degenerada y/o la base universal sea coincidente con la j -ésima posición de la i -ésima secuencia de sondeo, o la j -ésima posición del oligonucleótido de referencia antes de la introducción de las bases degeneradas y/o las bases universales coincida con la j -ésima posición del i -ésimo oligonucleótido, a_{ij} presenta un valor de coincidencia; en caso contrario, a_{ij} presenta un valor de no coincidencia. Particularmente, el valor de no coincidencia es 0 y el valor de coincidencia es un valor diferente de 0. Más particularmente, el valor de no coincidencia y el valor de coincidencia de a_{ij} son 0 y 1, respectivamente.

Además, $a_{ij} + d_j$ se selecciona como una variable binaria que representa si la secuencia de sondeo de referencia con o sin la base degenerada y/o base universal introducida coincide o no coincide en la j -ésima posición con la i -ésima secuencia de sondeo o el oligonucleótido de referencia con o sin la base degenerada y/o base universal introducida coincide o no coincide en la j -ésima posición con el i -ésimo oligonucleótido.

Particularmente, en el caso de que la base degenerada y/o la base universal se introduzca en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia, no se introduzca en la $(j+1)$ -ésima posición, y la j -ésima posición de la secuencia de sondeo de referencia sea coincidente con la j -ésima posición de la i -ésima secuencia de sondeo, o la j -ésima posición del oligonucleótido de referencia sea coincidente con la j -ésima posición del i -ésimo oligonucleótido, $a_{ij} + d_j$ presenta el valor de coincidencia; en caso contrario, $a_{ij} + d_j$ presenta el valor de no coincidencia. En el caso de que la $(j+1)$ -ésima posición de la secuencia de sondeo de referencia coincida con la $(j+1)$ -ésima posición de la i -ésima secuencia de sondeo, o la $(j+1)$ -ésima posición del oligonucleótido de referencia sea coincidente con la $(j+1)$ -ésima posición del i -ésimo oligonucleótido, $a_{i,j+1} + d_{j+1}$ presenta el valor de coincidencia; en caso contrario, $a_{i,j+1} + d_{j+1}$ presenta el valor de no coincidencia. Particularmente, el valor de no coincidencia es 0 y el valor de coincidencia es un valor diferente de 0. Más particularmente, el valor de no coincidencia es 0 y el valor de coincidencia es 1 o 2.

En la presente invención, el valor mínimo del conjunto $\{a_{ij} + d_j\}$ que presenta el valor de coincidencia y/o el valor de no coincidencia de $a_{ij} + d_j$ como elementos es mayor o igual a x_{ij} , una variable binaria que presenta un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para la selección o no de la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o del i -ésimo oligonucleótido de la pluralidad de oligonucleótidos. Por ejemplo, en el caso de que el valor mínimo del conjunto $\{a_{ij} + d_j\}$ que presenta el valor de coincidencia y/o el valor de no coincidencia de $a_{ij} + d_j$ como elementos presente un valor de no coincidencia, el valor mínimo de $\{a_{ij} + d_j\}$ debe ser igual a x_i con el valor de no selección ($x_{no\ sel.}$), y en el caso de que el valor mínimo del conjunto $\{a_{ij} + d_j\}$ que presenta el valor de coincidencia y/o el valor de no coincidencia de $a_{ij} + d_j$ como elementos presente un valor de coincidencia, el valor mínimo de $\{a_{ij} + d_j\}$ debe ser igual o superior a x_i con el valor de selección ($x_{sel.}$).

En otras palabras, el valor (valor de coincidencia) de $a_{ij} + d_j$ en la j -ésima posición de la i -ésima secuencia de sondeo o el i -ésimo oligonucleótido coincidente para la secuencia de sonda de referencia o el oligonucleótido de referencia en el que se introduce la base degenerada y/o base universal es igual o superior al valor seleccionado ($x_{sel.}$); el valor (valor de coincidencia) de $a_{ij} + d_j$ en la j -ésima posición de la i -ésima secuencia de sondeo o el i -ésimo oligonucleótido coincidente para la secuencia de sonda de referencia o el oligonucleótido de referencia en el que no se introduce la base degenerada y/o base universal es igual o mayor al valor seleccionado ($x_{sel.}$), y el valor (valor de no coincidencia) de $a_{ij} + d_j$ en la j -ésima posición de la i -ésima secuencia de sondeo o el i -ésimo oligonucleótido no coincidente para la secuencia de sonda de referencia o el oligonucleótido de referencia en el que no se ha introducido la base degenerada y/o base universal es igual o superior al valor de no selección ($x_{no\ sel.}$) y es inferior al valor de selección ($x_{sel.}$).

La fórmula de restricción 2 se ha establecido para cumplir dichas condiciones de restricción.

En el caso de que el método se lleve a cabo para conseguir la fórmula objetiva 1 y cumpla las fórmulas de restricción 1 y 2, pueden obtenerse los valores de las variables de decisión d_j y x_i como las soluciones óptimas, y la cobertura máxima del oligonucleótido preparado mediante la presente invención puede obtenerse mediante división de la suma de valores x_i por el número total de la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos.

Según una realización de la presente invención, el método selecciona las secuencias diana de ácido nucleico que van a detectarse y proporciona un oligonucleótido en el que se ha introducido por lo menos una de las bases degeneradas y/o bases universales en el número permisible predeterminado y se hibrida la secuencia diana seleccionada de ácido nucleico con ella.

Las secuencias diana de ácido nucleico que van a detectarse pueden seleccionarse a partir del valor x_i , que es la

solución óptima obtenida que cumple la fórmula objetiva 1 además de cumplir las fórmulas de restricción 1 y 2, y el oligonucleótido en el que se han introducido las bases degeneradas y/o las bases universales en el número permisible predeterminado para la hibridación con la secuencia diana seleccionada de ácido nucleico, mediante la solución óptima, el valor d_j .

Cabe destacar que mediante una fórmula objetiva puede seleccionarse no solo un oligonucleótido en el que se ha introducido por lo menos una base degenerada y/o base universal, sino también secuencias diana de ácido nucleico que serán la diana, es decir, que serán cubiertas por el oligonucleótido.

Según una realización, en donde la fórmula de restricción 1 son las fórmulas de restricción 1-1 y 1-2 a continuación:

Fórmula de restricción 1-1:

$$\sum_{j=1}^b d_j \leq D_{lim1}$$

Fórmula de restricción 1-2:

$$\sum_{j=1}^b d_j \leq D_{lim2}$$

en donde d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si se introduce la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o del oligonucleótido de referencia; j es un índice de posiciones de una secuencia de sondeo de referencia o un oligonucleótido de referencia comprendido entre 1 y c ; la secuencia de sondeo de referencia o el oligonucleótido de referencia comprende la primera posición comprendida entre 1 y b y la segunda parte está comprendida entre $b+1$ y c ; y D_{lim1} y D_{lim2} son números limitados de bases degeneradas y/o bases universales introducidas en la primera y segunda partes de la secuencia de sondeo de referencia o del oligonucleótido de referencia, respectivamente.

Mediante la modificación de la fórmula de restricción 1 en las fórmulas de restricción 1-1 y 1-2, puede llevarse a cabo la etapa (c) en la presente invención para conseguir la fórmula objetiva 1 siguiente, junto con el cumplimiento de las fórmulas de restricción 1-1, 1-2 y 2.

En la presente invención, puede aplicarse la programación lineal a dos partes diferentes de una secuencia de sondeo de referencia u oligonucleótido de referencia de manera independiente o simultánea, y particularmente puede aplicarse simultáneamente.

Debido a que las descripciones para la variable binaria d_j , el índice de posiciones j y el último índice c en las fórmulas de restricción 1-1 y 1-2 es idéntico a los de la fórmula de restricción 1, las descripciones comunes entre ellas se han omitido con el fin de evitar una innecesaria redundancia que lleve a hacer más compleja la presente especificación.

La secuencia de sondeo de referencia o el oligonucleótido de referencia comprende la primera parte comprendida entre 1 y b y la segunda parte comprendida entre $b+1$ y c .

El valor " b " es el índice de la última posición de la primera parte; $b+1$ es el índice de la primera posición de la segunda parte, y " c " es el índice de la última posición de la segunda parte.

Las longitudes de la primera parte y de la segunda parte pueden ser iguales o diferentes. En particular, la primera parte puede ser más larga que la segunda parte.

El valor " b " no se encuentra particularmente limitado y puede ser, por ejemplo, un número entero entre 15 y 40, entre 15 y 30 o entre 20 y 25.

D_{lim1} y D_{lim2} son números limitados de bases degeneradas y/o bases universales introducidas en la primera y segunda partes de la secuencia de sondeo de referencia o del oligonucleótido de referencia, respectivamente.

D_{lim1} y D_{lim2} pueden ser iguales o diferentes entre sí, o D_{lim1} puede ser mayor que D_{lim2} . Alternativamente, D_{lim} de la fórmula de restricción 1 puede ser igual o diferente de la suma de D_{lim1} y D_{lim2} .

El limitado número de bases degeneradas y/o bases universales introducidas en la primera parte de la secuencia de sondeo de referencia o del oligonucleótido de referencia, D_{lim1} , es particularmente 6, 4, 3 o 2. El limitado número de

bases degeneradas y/o bases universales introducidas en la segunda parte de la secuencia de sondeo de referencia o del oligonucleótido de referencia, D_{lim2} , es particularmente 5, 3, 2 o 1.

Según una realización del a presente invención, en la que la fórmula de restricción 1 son las fórmulas de restricción 1-1 y 1-3, a continuación:

Fórmula de restricción 1-1:

$$\sum_{j=1}^b d_j \leq D_{lim1}$$

Fórmula de restricción 1-3:

$$\sum_{j=b+n}^c d_j \leq D_{lim3}$$

en donde d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o en el oligonucleótido de referencia; j es un índice de posiciones de una secuencia de sondeo de referencia o de un oligonucleótido de referencia comprendido entre 1 y 2; la secuencia de sondeo de referencia o el oligonucleótido de referencia comprende una primera parte comprendida entre 1 y b , una segunda parte comprendida entre $b+1$ y $b+(n-1)$, y una tercera parte comprendida entre $b+n$ y c ; "n" es un número entero entre 2 y 10; D_{lim1} y D_{lim3} es un número limitado de bases degeneradas y/o bases universales introducidas en la primera y tercera parte de la secuencia de sondeo de referencia o del oligonucleótido de referencia, respectivamente, y la segunda parte representa una parte de separación que comprende bases universales, bases no naturales o bases no complementarias a las bases de la segunda parte.

Mediante la modificación de la fórmula de restricción 1 en las fórmulas de restricción 1-1 y 1-3, puede llevarse a cabo la etapa (c) en la presente invención a fin de conseguir la fórmula objetiva 1 siguiente, junto con el cumplimiento de las fórmulas de restricción 1-1, 1-3 y 2.

En la presente invención, la programación lineal puede aplicarse independientemente o simultáneamente a dos partes diferentes aparte de la parte de separación de entre las tres partes diferentes en una secuencia de sondeo de referencia u oligonucleótido de referencia, y particularmente, puede aplicarse en dos partes diferentes simultáneamente.

Debido a que las descripciones de una variable binaria d_j , un índice de las posiciones j , y el último índice de la primera parte b en las fórmulas de restricción 1-1 y 1-3 es idéntico a los de las fórmulas de restricción 1-1 y 1-2; las descripciones comunes entre ellas se han omitido a fin de evitar una innecesaria redundancia que conduciría a hacer más compleja la presente especificación.

La secuencia de sondeo de referencia o el oligonucleótido de referencia comprende la primera parte comprendida entre 1 y b , una segunda parte comprendida entre $b+1$ y $b+(n-1)$ y una tercera parte comprendida entre $b+n$ y c .

El valor "b" es el índice de la última posición de la primera parte; "b+1" es el índice de la primera posición de la segunda parte, "b+(n-1)" es el índice de la última posición de la segunda parte; "b+n" es el índice de la primera posición de la tercera parte, y "c" es el índice de la última posición de la tercera parte.

Además, "n" es un número entero entre 2 y 10, entre 3 y 10, entre 4 y 8 o entre 5 y 7.

Una secuencia de sondeo de referencia u oligonucleótido de referencia está separada en primera parte y tercera parte por la segunda parte como parte de separación, y las bases degeneradas y/o bases universales en un número permisible predeterminado se introducen en la primera parte y la tercera parte independiente o simultáneamente.

La segunda parte representa una parte de separación que comprende bases universales, bases no naturales o bases no complementarias a bases de la segunda parte.

La segunda parte como parte de separación comprende nucleótidos no contiguos o contiguos que presentan bases universales, bases no naturales o bases no complementarias a bases de la segunda parte.

El número de bases universales, bases no naturales o bases no complementarias a bases de la segunda parte

incluidas en la segunda parte como parte de separación es de entre 2 y 9, de entre 3 y 9, de entre 4 y 8 o de entre 5 y 7.

La descripción anterior de la base universal se aplica a la base universal incluida en la segunda parte.

La expresión utilizada en el presente documento "base no natural" se refiere a derivados de bases naturales, tales como adenina (A), guanina (G), timina (T), citosina (C) y uracilo (U), que son capaces de formar pares de bases de unidos mediante enlaces de hidrógeno. La expresión utilizada en el presente documento "base no natural" incluye bases que presentan patrones de apareamiento de bases diferentes de los de las bases naturales como compuestos madre, tal como se describe en, por ejemplo, las patentes U.S. n.º 5.432.272, n.º 5.965.364, n.º 6.001.983 y n.º 6.037.120. Entre los ejemplos específicos de bases no naturales se incluyen K, X, H, J, M, N, iso-C, iso-G, iso-dC e iso-dG.

Las bases no complementarias a bases de la segunda parte implican que, por ejemplo, en el caso de que las bases de las posiciones $j=b+1$ y $b+2$ en la segunda parte sean A y G, respectivamente, incluirán C o G y A o T, respectivamente.

Es crucial que la parte de separación en el oligonucleótido preparado mediante la presente invención presente la T_f mínima, en las tres partes, con el fin de que la parte de separación forme una estructura de burbuja de no apareamiento de bases bajo condiciones en las que la primera y segunda parte están apareadas con el ácido nucleico molde, permitiendo que una de la primera y segunda partes se diferencie de la otra parte en términos de especificidad de apareamiento con el ácido nucleico molde, de manera que la especificidad de apareamiento del oligonucleótido está determinado doblemente por la primera y segunda parte, con lo que la especificidad de apareamiento global del oligonucleótido resulta considerablemente mejorada.

Las longitudes de la primera parte y la tercera parte pueden ser iguales o diferentes. Particularmente, la primera parte puede ser más larga que la tercera parte.

D_{lim1} y D_{lim3} son los números limitados de bases degenerados y/o bases universales introducidos en la primera y tercera parte de la secuencia de sondeo de referencia o del oligonucleótido de referencia, respectivamente.

D_{lim1} y D_{lim3} pueden ser iguales o diferentes entre sí, o D_{lim1} puede ser mayor que D_{lim3} . Alternativamente, D_{lim} de la fórmula de restricción 1 puede ser igual o diferente de la suma de D_{lim1} y D_{lim3} .

El número limitado de bases degeneradas y/o de bases universales introducidas en la primera parte de la secuencia de sondeo de referencia o del oligonucleótido de referencia, D_{lim1} , es particularmente 6, 4, 3 o 2. El número limitado de bases degeneradas y/o bases universales introducidas en la tercera parte de la secuencia de sondeo de referencia o del oligonucleótido de referencia, D_{lim3} , es particular 5, 3, 2 o 1.

El término "cobertura" utilizado en el presente documento se refiere a que un oligonucleótido (un cebador o una sonda) es suficientemente complementario para hibridarse selectivamente con una secuencia diana de ácido nucleico bajo las condiciones de apareamiento designadas o bajo condiciones restrictivas, comprendiendo las expresiones "sustancialmente complementario" y "perfectamente complementario", particularmente "perfectamente complementario".

El término "hibridación" utilizado en el presente documento se refiere a formar un ácido nucleico de doble cadena a partir de un ácido nucleico de cadena sencilla complementario. Un oligonucleótido que se hibridará con una secuencia diana de ácido nucleico incluye no solo una secuencia que es perfectamente complementaria a una secuencia diana de ácido nucleico, sino también una secuencia que es suficiente para hibridarse específicamente con una secuencia diana de ácido nucleico bajo determinadas condiciones restrictivas. Por ejemplo, un oligonucleótido puede comprender uno o más nucleótidos no complementarios (es decir, no coincidencias) respecto a una secuencia diana de ácido nucleico, con la condición de que su especificidad no resulte perjudicada. Por lo tanto, en la presente invención, un oligonucleótido puede comprender una secuencia particularmente complementaria y una perfectamente complementaria respecto a una secuencia diana de ácido nucleico, y particularmente incluye una secuencia perfectamente complementaria (o una secuencia coincidente).

En el caso de que el método comprenda, además, entre las etapas (a) y (b), (a-1) la agrupación de la pluralidad de secuencias de sondeo o de la pluralidad de oligonucleótidos según la identidad de secuencia para obtener una pluralidad de patrones de secuencia, la etapa (c) de la presente invención puede llevarse a cabo de la manera siguiente.

Según una realización, la etapa (c) se lleva a cabo para conseguir la fórmula objetiva 2 siguiente junto con el cumplimiento de las fórmulas de restricción 3 y 4 siguientes:

Fórmula objetiva 2:

$$\text{Máx. : } \sum_{i=1}^r x_i$$

en donde Máx.: representa la maximización; p_i es el número de secuencias de sondeo pertenecientes al patrón de la i -ésima secuencia de las secuencias de sondeo o es el número de oligonucleótidos pertenecientes al patrón de la i -ésima secuencia de oligonucleótidos; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para la selección o no del patrón de la i -ésima secuencia de las secuencias de sondeo o el patrón de la i -ésima secuencia de oligonucleótidos, e "i" es un índice de patrones de secuencia de las secuencias de sondeo u oligonucleótidos comprendido entre 1 y r ;

Fórmula de restricción 3:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en la que d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; j es un índice de posiciones de la secuencia de sondeo de referencia o el oligonucleótido de referencia comprendido entre 1 y c , y D_{lim} es un número limitado de bases degeneradas y/o bases universales introducido en la secuencia de sondeo de referencia o del oligonucleótido de referencia;

Fórmula de restricción 4:

$$x_i \leq \min\{a_{i,j} + d_j\} \text{ para todo } i, j$$

en la que x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para la selección o no del patrón de la i -ésima secuencia de secuencias de sondeo o el patrón de la i -ésima secuencia de oligonucleótidos; $a_{i,j}$ es una constante binaria que representa si una secuencia de sondeo perteneciente al patrón de la i -ésima secuencia de secuencias de sondeo es coincidente o no coincidente en la j -ésima posición con la secuencia de sondeo de referencia o un oligonucleótido perteneciente al patrón de la i -ésima secuencia de oligonucleótidos es coincidente o no coincidente en la j -ésima posición con el oligonucleótido de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o la base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; $a_{i,j} + d_j$ es una variable binaria que representa si la secuencia de sondeo de referencia con o sin la base degenerada introducida y/o base universal es coincidente o no coincidente en la j -ésima posición con una secuencia de sondeo perteneciente al patrón de la i -ésima secuencia de secuencias de sondeo o el oligonucleótido de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con el oligonucleótido perteneciente al patrón de la i -ésima secuencia de oligonucleótidos; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min\{a_{i,j} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{i,j} + d_j\}$; y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo u oligonucleótidos.

Con el fin de aplicar la fórmula objetiva 1 y las fórmulas de restricción 1 y 2 a una pluralidad de patrones de secuencia de una pluralidad de secuencias de sondeo o de una pluralidad de oligonucleótidos, se reconstruye la fórmula objetiva 1, y las fórmulas de restricción 1 y 2 en la fórmula objetiva 2 y las fórmulas de restricción 3 y 4, para una pluralidad de patrones de secuencia. Por lo tanto, las descripciones de la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos en la fórmula objetiva 1, y las fórmulas de restricción 1 y 2 se representan mediante las descripciones de la pluralidad de patrones de secuencia en la fórmula objetiva 2, y las fórmulas de restricción 3 y 4. Además, las descripciones comunes entre la fórmula objetiva 1, las fórmulas de restricción 1 y 2, y la fórmula objetiva 2, y fórmulas de restricción 3 y 4, se han omitido a fin de evitar redundancias innecesarias que llevarían a una mayor complejidad de la presente especificación.

En la fórmula objetiva 2, p_i es el número de secuencias de sondeo pertenecientes al patrón de la i -ésima secuencia de secuencias de sondeo o el número de oligonucleótidos pertenecientes al patrón de la i -ésima secuencia de oligonucleótidos.

En donde el método se lleva a cabo para conseguir la fórmula objetiva 2 junto con el cumplimiento de las fórmulas de restricción 3 y 4, pueden obtenerse las variables de decisión d_j y x_i como soluciones óptimas.

En el caso de que se obtenga como una solución un valor de no selección ($x_{no\ sel.}$) o un valor de selección ($x_{sel.}$) del

patrón de la i -ésima secuencia de secuencias de sondeo o el patrón de la i -ésima secuencia de oligonucleótidos, puede obtenerse el número de secuencias diana de ácido nucleico detectadas mediante el oligonucleótido preparado mediante la presente invención mediante la multiplicación de x_i de cada patrón de secuencia por el número π_i de secuencias de sondeo u oligonucleótidos pertenecientes a cada patrón de secuencia y sumando el producto de π_i y x_i a la totalidad de los patrones de secuencia. Además, la cobertura diana máxima del oligonucleótido preparado mediante la presente invención puede determinarse mediante división del número de secuencias diana de ácido nucleico que va a detectarse por el número total de secuencias diana de ácido nucleico.

Según una realización, en donde la fórmula de restricción 3 son las fórmulas de restricción 3-1 y 3-2 a continuación:

Fórmula de restricción 3-1:

$$\sum_{j=1}^b d_j \leq D_{lim1}$$

Fórmula de restricción 3-2:

$$\sum_{j=b+1}^c d_j \leq D_{lim2}$$

en la que d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; j es un índice de posiciones de la secuencia de sondeo de referencia o el oligonucleótido de referencia comprendido entre 1 y c ; la secuencia de sondeo de referencia o el oligonucleótido de referencia comprende la primera parte comprendida entre 1 y b , y la segunda parte comprendida entre $b+1$ y c ; y D_{lim1} y D_{lim2} son números limitados de bases degeneradas y/o bases universales introducidas en la primera y segunda partes de la secuencia de sondeo de referencia o del oligonucleótido de referencia, respectivamente.

Según una realización de la presente invención, en donde la fórmula de restricción 3 son las fórmulas de restricción 3-1 y 3-3, a continuación:

Fórmula de restricción 3-1:

$$\sum_{j=1}^b d_j \leq D_{lim1}$$

Fórmula de restricción 3-3:

$$\sum_{j=b+n}^c d_j \leq D_{lim3}$$

en la que d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; j es un índice de posiciones de la secuencia de sondeo de referencia o el oligonucleótido de referencia comprendido entre 1 y c ; la secuencia de sondeo de referencia o el oligonucleótido de referencia comprende la primera parte comprendida entre 1 y b , y la segunda parte comprendida entre $b+1$ y $b+(n-1)$, y la tercera parte comprendida entre $b+n$ y c ; n es un número entero entre 2 y 10; D_{lim1} y D_{lim3} son números limitados de bases degeneradas y/o bases universales introducidas en la primera y tercera partes de la secuencia de sondeo de referencia o del oligonucleótido de referencia, respectivamente, y la segunda parte representa una parte de separación que comprende bases universales, bases no naturales o bases no complementarias a las bases de la segunda parte.

Debido a que las descripciones de las fórmulas de restricción 1-1 a 1-3 son idénticas a las de las fórmulas de restricción 3-1 a 3-3, se omiten las descripciones comunes entre ellas con el fin de evitar redundancia innecesaria que podría llevar a mayor complejidad de la presente especificación.

El método de la presente invención resulta particularmente útil en el caso de que se diseñe un oligonucleótido para detectar una pluralidad de patógenos con diversidad genética, tal como un virus, o para cribar géneros bacterianos (p. ej., *Campylobacter*, *Salmonella*, *Shigella*, *Vibrio* o *Aeromonas*).

II. Determinación de una región de sondeo en una pluralidad de secuencias diana de ácido nucleico.

En el segundo aspecto de la presente invención se proporciona un método según la reivindicación 7 para determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia, que comprende: (a) seleccionar una región de sondeo candidata en una pluralidad de secuencias diana de ácido nucleico y proporcionar secuencias en la región de sondeo candidata como una pluralidad de secuencias de sondeo, (b) introducir una base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre la pluralidad de secuencias de sondeo, de manera que la pluralidad de secuencias de sondeo sean coincidentes al máximo entre sí, y (c) determinar la región de sondeo candidata como región de sondeo para la detección de la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico en el caso de que la coincidencia máxima no sea inferior a una cobertura predeterminada.

Una región de sondeo se refiere a una región conservada que comprende secuencias que se han mantenido conservadoramente entre diferentes organismos, es decir, una secuencia conservada como una parte con la que se hibrida un cebador o una sonda. Una región conservada que es una parte biológicamente muy significativa representa una parte en la que las secuencias son similares o idénticas en diferentes moléculas de ácido nucleico en diferentes organismos. La región conservada se utiliza como un indicador muy importante para estudios filogenéticos y también se utiliza como un parte de sondeo en el caso de que se detecten diferentes organismos de una manera multiplex.

Según la presente invención, pueden determinarse de una manera única las secuencias mantenidas conservadoramente en diferentes organismos, y la región conservada que comprende dichas secuencias conservadas puede utilizarse como la parte (es decir, la región de sondeo) con la que se hibrida un cebador o una sonda.

La presente invención se describe como un método para determinar una región de sondeo, aunque ello también podría expresarse como un método para determinar una región conservada. En el caso de que se introduzca las bases degeneradas y/o las bases universales en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes en la pluralidad de secuencias de sondeo de acuerdo con la invención descrita anteriormente, de manera que la pluralidad de secuencias de sondeo sean coincidentes al máximo y la región de sondeo candidata puede presentarse como una región de sondeo en la que la coincidencia máxima no sea inferior a una cobertura predeterminada.

Debido a que el segundo método de la presente invención se basa en el primer método de la presente invención descrito anteriormente, las descripciones comunes entre ellos se han omitido a fin de evitar una redundancia innecesaria que conduciría a una mayor complejidad de la presente especificación.

La Fig. 3 es un diagrama de flujo de los procedimientos para llevar a cabo la invención de acuerdo con una realización de la misma. El método de la presente invención se describirá en referencia a la Fig. 3.

Etapas (a): provisión de una pluralidad de secuencias de sondeo (210).

En primer lugar, se selecciona una región de sondeo candidata en una pluralidad de secuencias diana de ácido nucleico y se proporcionan secuencias en la región de sondeo candidata como una pluralidad de secuencias de sondeo.

Según una realización, se alinea una pluralidad de secuencias diana de ácido nucleico previamente a la etapa (a).

Debido a que las descripciones de la secuencia diana de ácido nucleico y alineación en la etapa (a) del segundo método de la presente invención son idénticas a las del primer método de la presente invención, las descripciones comunes a ellas han sido omitidas a fin de evitar una redundancia innecesaria que podría llevar a una mayor complejidad de la presente especificación.

La pluralidad de secuencias de sondeo proporcionadas en la etapa (a) de la presente invención son una referencia de comparación y sujeto para determinar si hay una no coincidencia en la etapa (b), y son un sujeto de introducción para introducir bases degeneradas y/o bases universales para sus coincidencias máximas.

Según una realización de la presente invención, cada una de la pluralidad de secuencias de sondeo presenta un índice único.

Según una realización, la etapa (a) se lleva a cabo mediante la selección de por lo menos dos regiones de sondeo candidatas que están en diferentes ubicaciones alineadas con la pluralidad de secuencias diana de ácido nucleico, y proporcionar en cada una de las dos o más regiones de sondeo candidatas como una pluralidad de secuencias de sondeo.

Etapas (b): introducción de una base degenerada y/o base universal de manera que la pluralidad de secuencias de

sondeo sean coincidentes al máximo entre sí (220).

A continuación, se introduce una base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre la pluralidad de secuencias de sondeo, de manera que la pluralidad de secuencias de sondeo son coincidentes al máximo entre sí.

A diferencia del primer método e la presente invención, descrito anteriormente, la etapa (b) del segundo método de la presente invención puede llevarse a cabo sin proporcionar una secuencia de sondeo de referencia a modo de referencia para determinar si va a coincidir. En la etapa (b) del segundo método e la presente invención para determinar la región de sondeo, la pluralidad de secuencias de sondeo no solo es un sujeto de comparación, sino también un sujeto de introducción que introduce la base degenerada y/o base universal. Por ejemplo, en las secuencias de sondeo comprendidas entre 1 y 3, en el caso de que la base en la posición 3 de las secuencias de sondeo 1, 2 y 3 sea "A", "C" y "G", respectivamente, la base degenerada V (A, C o G) puede introducirse en la posición 3 para la máxima coincidencia.

La expresión "máxima coincidencia" utilizada en el presente documento en referencia a la introducción de la base degenerada y/o base universal significa que se incluyen las mismas secuencias tanto como resulte posible en una pluralidad de secuencias de sondeo mediante la introducción de la base degenerada y/o base universal, y la coincidencia máxima puede expresarse como una proporción o porcentaje con respecto al número total de la pluralidad de secuencias de sondeo.

Debido a que la descripción de la base degenerada y/o base universal en la etapa (b) del segundo método de la presente invención es igual a la del primer método de la presente invención, las descripciones comunes entre ellas han sido omitidas a fin de evitar una redundancia innecesaria que podría llevar a una mayor complejidad de la presente especificación.

Según una realización de la presente invención, la etapa (b) se lleva a cabo mediante la selección de una de la pluralidad de secuencias de sondeo a modo de secuencia de sondeo de referencia y la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la secuencia de sondeo de referencia en una o más posiciones no coincidentes entre la secuencia de sondeo de referencia y la pluralidad de secuencias de sondeo para la máxima coincidencia con la pluralidad de secuencias de sondeo.

En la etapa (b) de la presente invención, como referencia de comparación para determinar si hay una no coincidencia y un sujeto de introducción que introduce la base degenerada y/o base universal, en donde la secuencia de sondeo de referencia puede seleccionarse de una pluralidad de secuencias de sondeo.

Por ejemplo, en las secuencias de sondeo comprendidas entre 1 y 3, en el caso de que la base en la posición 3 de las secuencias de sondeo 1, 2 y 3 sea "A", "C" y "G", respectivamente, la secuencia de sondeo 1 se selecciona como la secuencia de sondeo de referencia, en comparación con las secuencias de sondeo 2 y 3, y después puede introducirse la base degenerada V (A, C o G) en la posición 3 de la secuencia de sondeo de referencia para la máxima coincidencia.

Según una realización, la secuencia de sondeo de referencia es una secuencia de sondeo que presenta el número más grande de secuencias idénticas de la pluralidad de secuencias de sondeo.

Según una realización de la presente invención, la etapa (b) se lleva a cabo mediante la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre sí en la pluralidad de patrones de secuencias para cada una de las dos o más regiones de sondeo candidatas, de manera que la pluralidad de patrones de secuencia son coincidentes al máximo entre sí.

Etapla (c): determinación de una región de sondeo (230).

Después, en el caso de que la coincidencia máxima no sea inferior a una cobertura predeterminada, la región de sondeo candidata se determina como una región de sondeo para detectar la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico.

El término "cobertura" utilizado en el presente documento en referencia a la región de sondeo en la presente invención se refiere a la proporción de secuencias perfectamente coincidentes en la pluralidad de secuencias de sondeo en las que se ha introducido la base degenerada y/o base universal. La cobertura puede expresarse mediante una proporción o en porcentaje.

La cobertura es de 20 % o superior, de 40 % o superior, de 60 % o superior, de 80 % o superior, de 90 % o superior, o de 100 % del número total de la pluralidad de secuencias de sondeo.

En el caso de que la pluralidad de secuencias de sondeo sean coincidentes al máximo entre sí mediante la introducción de la base degenerada y/o base universal, y la coincidencia máxima no sea inferior a una cobertura predeterminada,

la región de sondeo candidata en la que se determina la pluralidad de secuencias de sondeo se determina como la región de sondeo para detectar la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico.

Según una realización, la etapa (c) se lleva a cabo mediante la determinación como región de sondeo para detectar la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico, entre las dos o más regiones de sondeo candidatas, (i) una región de sondeo candidata en la que la coincidencia máxima no sea inferior a la cobertura predeterminada y se muestra la coincidencia máxima más alta, (ii) una región de sondeo candidata en la que la coincidencia máxima es no inferior a la cobertura predeterminada y se ha introducido un número mínimo de la base degenerada y/o base universal para la máxima coincidencia, o (iii) una región de sondeo candidata en la que la coincidencia máxima es no inferior a la cobertura predeterminada, en donde se muestra la coincidencia máxima más alta y se ha introducido un número mínimo de la base degenerada y/o base universal para la máxima coincidencia.

Por ejemplo, las regiones de sondeo candidatas 1 a 3, al encontrarse en ubicaciones diferentes, se seleccionan en la alineación de la pluralidad de secuencias diana de ácido nucleico, la cobertura predeterminada es de 90 % o superior, y se introducen tres o menos bases degeneradas en la pluralidad de secuencias de sondeo para cada una de las regiones de sondeo candidatas 1 a 3 en una o más posiciones no coincidentes en la pluralidad de secuencias de sondeo. Como resultado, la región de sondeo candidata 1 representa una tasa de coincidencia máxima de 95 % mediante la introducción de una base degenerada, en donde la región de sondeo candidata 2 presenta una tasa de coincidencia máxima de 99 % mediante la introducción de dos bases degeneradas, y en donde la región de sondeo candidata 3 presenta una tasa de coincidencia máxima de 99 % mediante la introducción de tres bases degeneradas.

Según el criterio (i) para determinar la región de sondeo, las regiones de sondeo candidatas 2 y 3, como regiones de sondeo pueden determinarse aquellas que presentan una cobertura no inferior a 90 % y una tasa de coincidencia máxima más alta igual a 99 %, debido a que puede introducirse la base degenerada en la tres. Alternativamente, según el criterio (i), aunque la tasa de coincidencia máxima es no inferior a la cobertura predeterminada en todas las regiones de sondeo candidatas 1 a 3, puede determinarse la región de sondeo candidata 1 como la región de sondeo debido a que se utiliza el número más pequeño de bases degeneradas, es decir, una base degenerada, para la máxima coincidencia. Alternativamente, según el criterio (iii), como la región de sondeo, puede determinarse la región de sondeo candidata 2, en la que la tasa de coincidencia máxima es no inferior a la cobertura predeterminada, la coincidencia máxima es la más alta y se introduce el número mínimo de bases degeneradas para la coincidencia máxima más alta.

Según una realización de la presente invención, el método comprende, además, entre las etapas (a) y (b), (a-1) la agrupación de la pluralidad de secuencias de sondeo según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, en donde la etapa (b) se lleva a cabo mediante la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre sí en la pluralidad de patrones de secuencia, de manera que la pluralidad de patrones de secuencia sean coincidentes al máximo entre sí, y la etapa (c) se lleva a cabo mediante la determinación de la región de sondeo candidata a la que pertenece el patrón de secuencia como región de sondeo para detectar la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico en el caso de que la coincidencia máxima sea no inferior a una cobertura predeterminada.

Según una realización, el método comprende, además, entre las etapas (a) y (b), (a-1) la agrupación de la pluralidad de secuencias de sondeo según identidad de secuencia de cada una de las dos o más regiones de sondeo candidatas según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, en donde la etapa (b) se lleva a cabo mediante la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre sí en la pluralidad de patrones de secuencia para cada una de las dos o más regiones de sondeo candidatas, de manera que la pluralidad de patrones de secuencia sean coincidentes al máximo entre sí, y la etapa (c) se lleva a cabo mediante la determinación de la región de sondeo para detectar la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico entre las dos o más regiones de sondeo candidatas, (i) una región de sondeo candidata de un patrón de secuencia en la que la coincidencia máxima es no inferior a la cobertura predeterminada y se muestra la coincidencia máxima más alta, (ii) una región de sondeo candidata de un patrón de secuencia en la que la coincidencia máxima es no inferior a la cobertura predeterminada y se introduce un número mínimo de bases degeneradas y/o bases universales para la coincidencia máxima, o (iii) una región de sondeo candidata de un patrón de secuencia en el que la coincidencia máxima es no inferior a la cobertura predeterminada, se muestra la correspondencia máxima más alta y se introduce un número mínimo de la base degenerada y/o base universal para la coincidencia máxima.

Mediante la agrupación de la pluralidad de secuencias de sondeo según identidad de secuencia para obtener una pluralidad de patrones de secuencia, resulta posible reducir el número de secuencias de sondeo como sujeto de comparación para determinar si va a ser coincidente o no coincidente, reduciendo de esta manera el tiempo para determinar la región de sondeo.

Las etapas anteriormente descritas (b) y (c) también se aplican a una pluralidad de patrones de secuencia de una pluralidad de secuencias de sondeo.

Según una realización de la presente invención, cada una de la pluralidad de patrones de secuencia presenta un número de secuencia único.

5 Según una realización, la secuencia de sondeo de referencia es una secuencia de sondeo de un patrón de secuencia que presenta el número más grande secuencias de sondeo entre la pluralidad de patrones de secuencia.

10 La otra característica de la presente invención es la aplicación de programación lineal como lógica de optimización, en la determinación como región de sondeo mediante la introducción óptima de una base degenerada y/o base universal en un número permisible predeterminado en una pluralidad de secuencias de sondeo en una región de sondeo candidata.

15 Debido a que la descripción de una programación lineal en el segundo método de la presente invención es la misma que del primer método de la presente invención, las descripciones comunes a los dos se han omitido a fin de evitar una redundancia innecesaria que podría llevar a una mayor complejidad de la presente especificación.

Según la presente invención, la etapa (b) se lleva a cabo para conseguir la fórmula objetiva 3 siguiente junto con el cumplimiento de las fórmulas de restricción 5 y 6, a continuación:

20 Fórmula de restricción 3:

$$\text{Máx. : } \sum_{i=1}^r x_i$$

25 en donde Max.: representa la maximización; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo, e "i" es un índice de secuencias de sondeo comprendidas entre 1 y r,

Fórmula de restricción 5:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

30 en la que d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la pluralidad de secuencias de sondeo o la secuencia de sondeo de referencia; j es un índice de posiciones de la secuencia de sondeo de referencia o el oligonucleótido de referencia comprendido entre 1 y c; y D_{lim} es un número limitado de la base degenerada y/o base universal introducida en la pluralidad de secuencias de sondeo o la secuencia de sondeo de referencia;

Fórmula de restricción 6:

40 $x_i \leq \min\{a_{i,j} + d_j\}$ para todo i, j

45 en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo; $a_{i,j}$ es una constante binaria que representa si la pluralidad de secuencias de sondeo son coincidentes o no coincidentes entre sí en la posición j -ésima o si la i -ésima secuencia de sondeo es coincidente o no coincidente en la j -ésima posición con la secuencia de sondeo de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la pluralidad de secuencias de sondeo o la secuencia de sondeo de referencia; $a_{i,j} + d_j$ es una variable binaria que representa si la pluralidad de secuencias de consenso con o sin la base degenerada y/o base universal introducida son coincidentes o no coincidentes entre sí en la j -ésima posición o la secuencia de sondeo de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con la i -ésima secuencia de sondeo; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min\{a_{i,j} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{i,j} + d_j\}$ y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo.

55 Con el fin de aplicar la fórmula objetiva 1, las fórmulas de restricción 1 y 2 a un método para determinar una región de sondeo, se reconstruyó la fórmula objetiva 1, y las fórmulas de restricción 1 y 2, en la fórmula objetiva 3 y las fórmulas de restricción 5 y 6, para una pluralidad de secuencias de sondeo o una secuencia de sondeo de referencia, que es una referencia de comparación y sujeto de una no coincidencia, y un sujeto de introducción de la base degenerada

y/o base universal. Por lo tanto, la fórmula objetiva 1, y las fórmulas de restricción 1 y 2 se modifican para la aplicación a una pluralidad de secuencias de consenso o una secuencia de sondeo de referencia, que es una referencia de comparación y sujeto de una no coincidencia, y un sujeto de introducción de la base degenerada y/o base universal, proporcionando la fórmula objetiva 3, y las fórmulas de restricción 5 y 6. Además, las descripciones comunes a la fórmula objetiva 1, y fórmulas de restricción 1 y 2, y la fórmula objetiva 3, y fórmulas de restricción 5 y 6, se han omitido a fin de evitar una redundancia innecesaria que podría llevar a una mayor complejidad de la presente especificación.

En el caso de que el método comprenda, además, entre las etapas (a) y (b), (a-1) la agrupación de la pluralidad de secuencias de sondeo según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, o el método comprende, además, entre las etapas (a) y (b), (a-1) la agrupación de la pluralidad de secuencias de sondeo de cada una de las dos o más regiones de sondeo candidatas según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, puede llevarse a cabo la etapa (b) de la presente invención del modo siguiente.

Según una realización, la etapa (b) se lleva a cabo para conseguir la fórmula objetiva 4 siguiente, junto con el cumplimiento de las fórmulas de restricción 7 y 8, siguientes:

Fórmula objetiva 4:

$$\text{Máx.: } \sum_{i=1}^r p_i x_i$$

en donde Max: representa maximización; p_i es el número de secuencias de sondeo pertenecientes al patrón de la i -ésima secuencia de secuencias de sondeo; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no el patrón de la i -ésima secuencia de las secuencias de sondeo, e " r " es un índice de patrones de secuencia de las secuencias de sondeo comprendido entre 1 y r ,

Fórmula de restricción 7:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en donde d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo del patrón de secuencia o la secuencia de sondeo de referencia; j es un índice de posiciones de la secuencia de sondeo del patrón de secuencia o la secuencia de sondeo de referencia comprendido entre 1 y c ; y D_{lim} es un número limitado de la base degenerada y/o base universal introducida en la secuencia de sondeo del patrón de secuencia o la secuencia de sondeo de referencia;

Fórmula de restricción 8:

$$x_i \leq \min\{a_{i,j} + d_j\} \text{ para todo } i, j$$

en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no el patrón de la i -ésima secuencia de secuencias de sondeo; $a_{i,j}$ es una constante binaria que representa si las secuencias de sondeo de cada uno de los patrones de secuencia son coincidentes o no coincidentes entre sí en la j -ésima posición o una secuencia de sondeo del patrón de la i -ésima secuencia es coincidente o no coincidente en la j -ésima posición con la secuencia de sondeo de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de cada uno de los patrones de secuencia o la secuencia de sondeo de referencia; $a_{i,j} + d_j$ es una variable binaria que representa si las secuencias de sondeo de cada uno de los patrones de secuencia con o sin la base degenerada y/o base universal introducida son coincidentes o no coincidentes entre sí en la j -ésima posición o la secuencia de sondeo de referencia con o sin la base degenerada y/o base universal introducida son coincidentes o no coincidentes en la j -ésima posición con la secuencia de consenso del patrón de la i -ésima secuencia; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min\{a_{i,j} + d_j\}$ representa un valor mínimo de los elementos del conjunto $\{a_{i,j} + d_j\}$ y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo.

Con el fin de aplicar la fórmula objetiva 3, y las fórmulas de restricción 5 y 6, a una pluralidad de patrones de secuencia de una pluralidad de secuencias de sondeo, se reconstruyeron la fórmula objetiva 3, y las fórmulas de restricción 5 y 6, en la fórmula objetiva 4, y las fórmulas de restricción 7 y 8, para una pluralidad de patrones de secuencia. Por lo

tanto, las variables y constantes para la pluralidad de secuencias de consenso en la fórmula objetiva 3, y en las fórmulas de restricción 5 y 6, se modifican para la pluralidad de patrones de secuencia en la fórmula objetiva 4, y en las fórmulas de restricción 7 y 8. Además, las descripciones comunes a la fórmula objetiva 3, y fórmulas de restricción 5 y 6, y la fórmula objetiva 4, y fórmulas de restricción 7 y 8, se han omitido a fin de evitar una redundancia innecesaria que podría llevar a una mayor complejidad de la presente especificación.

III. Medio de almacenamiento, dispositivo y programa.

En otro aspecto de la presente invención, se proporciona un medio de almacenamiento legible por ordenador que contiene instrucciones para configurar un procesador para llevar a cabo un método de preparación de un oligonucleótido para la detección de una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima, en donde el método comprende: (a) seleccionar cada una de las secuencias con similitud de secuencia en la misma región de una pluralidad de secuencias diana de ácido nucleico como oligonucleótido para preparar una pluralidad de oligonucleótidos, (b) seleccionar uno de la pluralidad de oligonucleótidos como oligonucleótido de referencia, y (c) introducir una base degenerada y/o base universal en un número permisible predeterminado en el oligonucleótido de referencia en una o más posiciones no coincidentes entre el oligonucleótido de referencia y la pluralidad de oligonucleótidos para la coincidencia máxima con la pluralidad de oligonucleótidos, y seleccionar como una secuencia de un oligonucleótido, una secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una de base degenerada y/o base universal, preparando de esta manera un oligonucleótido con una cobertura diana máxima.

En todavía otro aspecto de la presente invención, se proporciona un programa informático para el almacenamiento en un medio de almacenamiento legible por ordenador, para configurar un procesador para llevar a cabo un método de preparación de un oligonucleótido para la detección de una pluralidad de secuencias diana de ácido nucleico que presenta similitud de secuencia con una cobertura diana máxima, y el método es el mismo que el método descrito en el medio de almacenamiento legible por ordenador descrito anteriormente.

En otro aspecto de la presente invención, se proporciona un dispositivo de preparación de un oligonucleótido para detectar una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima, que comprende (a) un procesador de ordenador, y (b) un medio de almacenamiento legible por ordenador del presente método acoplado al procesador de ordenador.

En otro aspecto de la presente invención, se proporciona un medio de almacenamiento legible por ordenador que contiene instrucciones para configurar un procesador para llevar a cabo un método de determinación de una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia, en donde el método comprende: (a) seleccionar una región de sondeo candidata en una pluralidad de secuencias diana de ácido nucleico y proporcionar secuencias en la región de sondeo candidata como pluralidad de secuencias de sondeo, (b) introducir una base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre la pluralidad de secuencias de sondeo, de manera que la pluralidad de secuencias de sondeo son coincidentes al máximo entre sí, y (c) determinar la región de sondeo candidata como una región de sondeo para la detección de la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico en el caso de que la coincidencia máxima es no inferior a una cobertura predeterminada.

En todavía otro aspecto de la presente invención, se proporciona un programa informático para el almacenamiento en un medio de almacenamiento legible por ordenador, para configurar un procesador para llevar a cabo un método de determinación de una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia, y el método es el mismo que el método descrito en el medio de almacenamiento legible por ordenador descrito anteriormente.

En todavía otro aspecto de la presente invención, se proporciona un dispositivo para determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presenta similitud de secuencia, que comprende: (a) un procesador informático, y (b) un medio de almacenamiento legible por ordenador del presente método acoplado con el procesador informático.

Debido a que el medio de almacenamiento, el dispositivo y el programa informático de la presente invención están destinados a llevar a cabo los presentes métodos descritos anteriormente en el presente documento en un ordenador, se han omitido las descripciones comunes a ellos a fin de evitar una redundancia innecesaria que podría llevar a una mayor complejidad de la presente especificación.

Las instrucciones del programa son operativas, al ser ejecutadas por el procesador, causando que el procesador lleve a cabo el presente método, descrito anteriormente. Las instrucciones de programa para llevar a cabo el método de preparación de un oligonucleótido descrito anteriormente puede comprender las instrucciones siguientes: (i) una instrucción para seleccionar cada una de las secuencias con similitud de secuencia en la misma región de sondeo de una pluralidad de secuencias diana de ácido nucleico como secuencia de sondeo o un oligonucleótido a fin de preparar una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos, (ii) una instrucción para seleccionar una

de la pluralidad de secuencias de sondeo como secuencia de sondeo de referencia o uno de la pluralidad de oligonucleótidos a modo de oligonucleótido de referencia, (iii) una instrucción para introducir una base degenerada y/o base universal en un número permisible predeterminado en la secuencia de consenso de referencia o en el oligonucleótido de referencia en una o más posiciones no coincidentes entre la secuencia de sondeo de referencia o el oligonucleótido de referencia y la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos para la coincidencia máxima con la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos, (iv) una instrucción para seleccionar como secuencia de un oligonucleótido, la secuencia de sondeo de referencia o una secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una de la base degenerada y/o la base universal, preparando de esta manera un oligonucleótido con una cobertura diana máxima. Además, las instrucciones del programa pueden incluir por lo menos una de las instrucciones siguientes: (v) una instrucción para recoger una pluralidad de secuencias diana de ácido nucleico a partir de una base de datos de secuencias de ácido nucleico, y (vi) una instrucción para alinear una pluralidad de secuencias diana de ácido nucleico.

El medio de almacenamiento y el dispositivo de la presente invención pueden incluir una base de datos de secuencias de ácido nucleico.

El presente método se implementa en un procesador, tal como un procesador en un ordenador autónomo o en un ordenador conectado a una red.

Entre los tipos de medio de almacenamiento legible por ordenador se incluyen diversos medios de almacenamiento, tales como CD-R, CD-ROM, DVD, memoria flash, disquete, disco duro, HDD portátil, USB, cinta magnética, MINIDISC, tarjeta de memoria no volátil, EEPROM, disco óptico, medio de almacenamiento óptico, RAM, ROM, memoria de sistema y servidor web.

El oligonucleótido, secuencias diana de ácido nucleico y/o una región de sondeo preparados en la presente invención pueden proporcionarse en una variedad de formas. Por ejemplo, el oligonucleótido preparado, las secuencias diana de ácido nucleico y/o una región de sondeo pueden proporcionarse a un sistema separado, tal como un sistema informático de sobremesa mediante una conexión de red (p. ej., LAN, VPN, intranet e internet) o conexión directa (p. ej., USB u otra conexión directa, por cable o inalámbrica), o proporcionarse en un medio portátil, tal como un CD, DVD, disquete y HDD portátil. De manera similar, el oligonucleótido preparado, secuencias diana de ácido nucleico y/o una región de sondeo pueden proporcionarse a un sistema de servidores mediante una conexión de red (p. ej., LAN, VPN, internet, intranet y red de comunicación inalámbrica) a un cliente, tal como un portátil o un sistema de ordenador de sobremesa.

Las instrucciones para configurar el procesador para llevar a cabo la presente invención pueden estar incluidas en un sistema lógico. Las instrucciones pueden descargarse y almacenarse en un módulo de memoria (p. ej., un disco duro u otra memoria, tal como una RAM o ROM local o conectada), aunque las instrucciones pueden proporcionarse en cualquier medio de almacenamiento de software, tal como un HDD portátil, USB, disquete, CD y DVD. Un código informático para implementar la presente invención puede implementarse en una variedad de lenguajes de programación, tales como C, C++, Java, Visual Basic, VBScript, JavaScript, Perl y XML. Además, puede utilizarse una variedad de lenguajes y protocolos en almacenamiento externo e interno y transmisión de datos y comandos según la presente invención.

El procesador informático puede prepararse de tal manera que un único procesador puede llevar a cabo varias acciones. Alternativamente, la unidad de procesador puede prepararse de tal manera que varios procesadores lleven a cabo varias acciones, respectivamente.

Las características y ventajas de la presente invención se resumen del modo siguiente:

(a) Diseño de un oligonucleótido para detectar una pluralidad de secuencias de ácido nucleico de una molécula diana de ácido nucleico que muestra diversidad genética con una cobertura diana máxima, la presente invención proporciona un método más lógico y eficiente mediante la adopción de una estrategia de selección de una secuencia de sondeo de referencia o un oligonucleótido de referencia para preparar un oligonucleótido en el que se introducen bases degeneradas y/o bases universales. Según una realización de la presente invención, la presente invención prepara un oligonucleótido que presenta una cobertura diana máxima en la que se introduce un número apropiado de bases degeneradas *in silico*, mediante la conversión de una lógica adoptada en la presente invención en una lógica matemática y que se programa informáticamente.

(b) La presente invención considera tanto la cobertura de la diana como la eficiencia de los oligonucleótidos en la preparación de oligonucleótidos en los que se introducen las bases degeneradas y/o las bases universales. Los oligonucleótidos con bases degeneradas introducidas incrementan la cobertura de la diana, aunque adolecen del problema de reducir la eficiencia de los oligonucleótidos, en particular causando una gran interferencia entre oligonucleótidos en la detección multiplex. Por lo tanto, el oligonucleótido debe prepararse considerando tanto la cobertura de la diana como el número de bases degeneradas que va a introducirse, y este objetivo técnico es conseguido por la presente invención. Según la presente invención, puede proporcionarse un oligonucleótido óptico de una manera más lógica y eficiente en términos de cobertura de la diana y eficiencia del oligonucleótido.

(c) Al introducir bases degeneradas y/o bases universales en un oligonucleótido con el fin de detectar una pluralidad de secuencias de ácido nucleico de una molécula diana de ácido nucleico que muestra diversidad genética con una cobertura diana máxima, se han utilizado convencionalmente métodos empíricos y manuales, que son un proceso que requiere mucho tiempo y es laborioso, de velocidad y exactitud reducidas.

Según una realización, a diferencia de los métodos convencionales descritos anteriormente, la presente invención puede presentar una introducción óptima de bases degeneradas y/o bases universales en un oligonucleótido de una manera lógica y automática. Al presentar una introducción óptima de bases degeneradas y/o bases universales en un oligonucleótido, el método de la presente invención resulta mucho más rápido y exacto que cualquier método convencional.

(d) Según una realización de la presente invención, la lógica de optimización de la presente invención puede utilizarse en (i) la preparación de un oligonucleótido en el que se introduce un número limitado de bases degeneradas y/o bases universales para la detección de una pluralidad de secuencias diana de ácido nucleico con una cobertura diana máxima, y (ii) la determinación de una región de sondeo en una pluralidad de secuencias diana de ácido nucleico.

A continuación, se describirá la presente invención en mayor detalle mediante ejemplos. Resultará evidente para el experto en la materia que estos ejemplos se pretende que sean más concretamente ilustrativos y que el alcance de la presente invención tal como se expone en las reivindicaciones adjuntas no se encuentra limitado por o mediante los ejemplos.

Ejemplos

Ejemplo 1: alineación de secuencias diana de ácido nucleico y clasificación (agrupación) de patrones de secuencia.

Las secuencias de UTR 5' de *Enterovirus* como pluralidad de secuencias diana de ácido nucleico recogidas del National Center for Biotechnology Information (NCBI) se alinearon y las secuencias recogidas se clasificaron en patrones de secuencia según la identidad de secuencia. Como resultado, en un total de 531 secuencias diana de ácido nucleico, se clasificaron en total 20 patrones de secuencia para una región que mostraba similitud de secuencia, y de entre ellas, la secuencia del vigésimo patrón de secuencia que contenía el número más alto de secuencias idénticas se seleccionó como una secuencia de ácido nucleico de referencia (es decir, un oligonucleótido de referencia). Los patrones de secuencia que contenían secuencias con tres o menos variaciones en comparación con la secuencia de ácido nucleico de referencia se resumen en la Tabla 1, a continuación.

Tabla 1

<i>i</i>	Secuencia de oligonucleótido																			<i>p_i</i>
1	C	A	T	C	A	G	G	C	C	C	C	T	G	A	A	T	G	C	G	1
2	C	T	T	C	C	G	T	C	C	C	C	T	G	A	A	T	G	C	G	1
3	C	G	T	C	C	G	G	C	C	C	C	T	G	A	A	T	G	C	G	3
4	C	C	A	C	C	G	G	C	C	C	C	T	G	A	A	T	G	C	G	1
5	C	C	A	C	C	G	G	C	C	C	C	T	G	A	A	T	G	C	G	1
6	C	C	T	G	C	G	G	C	C	C	C	T	G	A	A	T	G	C	G	1
7	C	C	T	C	G	G	G	A	C	C	C	T	G	A	A	T	G	C	G	2
8	C	C	T	C	C	A	G	C	C	C	C	T	G	A	A	T	G	C	G	1
9	C	C	T	C	C	G	T	G	C	T	C	T	G	A	A	T	G	C	G	5
10	C	C	T	C	C	G	G	C	C	C	C	T	G	A	A	T	G	C	T	1
11	C	C	T	C	C	G	G	A	C	C	C	T	G	A	A	T	A	C	G	3
12	C	C	T	C	C	G	G	C	A	C	C	T	G	A	A	T	G	G	G	2
13	C	C	T	C	C	G	G	C	C	C	C	A	G	A	A	T	G	A	G	2
14	C	C	T	C	C	G	G	C	C	C	C	C	G	A	A	T	G	C	G	1
15	C	C	T	C	C	G	G	C	C	C	C	T	G	A	A	A	G	C	G	1
16	C	C	T	C	C	G	G	C	C	C	C	T	G	A	A	T	G	G	A	1
17	C	C	T	C	C	G	G	C	C	C	C	T	G	A	A	T	G	G	G	1
18	C	C	T	C	C	G	G	C	C	C	C	T	G	A	A	T	G	A	G	86
19	C	C	T	C	C	G	G	C	C	C	C	T	G	A	A	T	G	C	T	4
20	C	C	T	C	C	G	G	C	C	C	C	T	G	A	A	T	G	C	G	413

El valor "i" en la Tabla 1 es un índice de patrón de secuencia y las secuencias de los patrones de secuencia son SEQ ID n.º 1 a n.º 20, respectivamente. Los números (*p_i*) en la Tabla 1 representan el número de secuencias diana de ácido nucleico pertenecientes a cada patrón de secuencia. Los patrones de secuencia en la Tabla 1 sirven tanto de secuencias diana de ácido nucleico que van a detectarse como de secuencias oligonucleotídicas (cebadores y/o sondas). Por ejemplo, en el caso de que se diseñe una sonda mediante la utilización de una secuencia de ácido nucleico del patrón de la vigésima secuencia de la Tabla 1, puede utilizarse para detectar una secuencia de ácido nucleico que contiene una secuencia complementaria a la secuencia de ácido nucleico del patrón de la vigésima

secuencia. En el caso de que se diseñe una sonda mediante la utilización de una secuencia complementaria a la secuencia de ácido nucleico del patrón de la vigésima secuencia, puede utilizarse para detectar una secuencia de ácido nucleico que contiene la secuencia de ácido nucleico del patrón de la vigésima secuencia. En los Ejemplos siguientes, se representan sondas mediante las secuencias en la Tabla 1, particularmente para introducir bases degeneradas (una sonda de referencia: una sonda de patrón de la vigésima secuencia, número total de sondas de 20 patrones de secuencia: 531).

Ejemplo 2: introducción de un número limitado (tres) de bases degeneradas en una sonda para coincidir al máximo entre sondas utilizando un algoritmo de programación lineal.

En el caso de que se utilice una sonda para detectar la pluralidad de secuencias diana de ácido nucleico con una cobertura diana máxima, se aplica un algoritmo de programación lineal para proporcionar una lógica de optimización adecuada para la aplicación óptima de bases degeneradas a la sonda.

Se utilizaron los veinte (20) patrones de secuencia mostrados en la Tabla 1. Debido a que una sonda a proporcionar debe detectar secuencias diana de ácido nucleico con una cobertura diana máxima e introducir bases degeneradas en posiciones de no coincidencia de la sonda, se establecieron x_i y d_j como una variable de decisión. En este caso, x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}=0$) y un valor de selección ($x_{sel.}=1$) para si se selecciona o no el patrón de la i -ésima secuencia de sondas; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}=0$) y un valor de introducción ($d_{int.}=1$) que representan la posibilidad de introducir una base degeneración en la j -ésima posición de la sonda de referencia (la secuencia de ácido nucleico de referencia del Ejemplo 1 se proporciona como la sonda de referencia). El valor "i" es un índice de patrones de secuencia de sondas comprendido entre 1 y 20; "j" es un índice de la posición de la sonda de referencia comprendido entre 1 y 20.

Una función objetiva es la maximización del número de sondas pertenecientes a los patrones de secuencia de sonda seleccionados (es decir,

$$Z=1x_1+1x_2+3x_3+1x_4+1x_5+1x_6+2x_7+1x_8+5x_9+1x_{10}+3x_{11}+2x_{12}+2x_{13}+1x_{14}+1x_{15}+1x_{16}+1x_{17}+86x_{18}+4x_{19}+413x_{20}).$$

Debido a que se pretende la introducción de un número limitado (tres) de bases degeneradas en la sonda, puede expresarse una fórmula de restricción como

$$d_1+d_2+d_3+d_4+d_5+d_6+d_7+d_8+d_9+d_{10}+d_{11}+d_{12}+d_{13}+d_{14}+d_{15}+d_{16}+d_{17}+d_{18}+d_{19}+d_{20} \leq 3.$$

Con el fin de utilizar el algoritmo de programación lineal considerando el patrón de coincidencia de la secuencia de sonda del Ejemplo 1, se adoptó la sonda del patrón de la vigésima secuencia en las secuencias de sonda de la Tabla 1 a modo de sonda de referencia. a_{ij} es 1 en el caso de que cada una de las secuencias de sonda sea coincidente con la sonda de referencia y a_{ij} es 0 en el caso de que no sea coincidente. Se muestran los resultados en la fig. 1 como una matriz $A=\{a_{ij}\}$.

Tal como se muestra en la Fig. 1, en el caso de que el patrón de secuencia de la sonda sea perfectamente coincidente con la sonda de referencia, tal como el elemento de la vigésima línea ($i=20$), a_{ij} es 1 para todo j , se ha encontrado que un valor mínimo de a_{ij} es 1.

Además, para la coincidencia máxima entre la sonda de referencia y los patrones de secuencia de sonda, se requiere una fórmula de restricción relacionada con la posibilidad de que la sonda de referencia sea coincidente o no con los patrones de secuencia de sonda antes y después de introducir bases degeneradas en la sonda de referencia. Los patrones 17°, 18°, 19° y 20° ($i=17, 18, 19, 20$) en la Fig. 1 se resumen en la Tabla 2 para la descripción de la fórmula de restricción.

Tabla 2

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
$a_{17,j}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	No seleccionable
$a_{18,j}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	No seleccionable
$a_{19,j}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	No seleccionable
$a_{20,j}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	Seleccionable

d_j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
-------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
$a_{17,j} + d_j$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	Seleccionable
$a_{18,j} + d_j$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	Seleccionable
$a_{19,j} + d_j$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	0	2	No seleccionable
$a_{20,j} + d_j$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	Seleccionable

5 En la expresión de si la sonda de referencia es coincidente o no con una sonda previamente a la introducción de bases degeneradas, a_{ij} es una constante binaria que representa si una sonda perteneciente al patrón de la i -ésima secuencia de sonda es coincidente o no coincidente en la j -ésima posición con la sonda de referencia, y un valor de no coincidencia y un valor de coincidencia de a_{ij} son 0 y 1, respectivamente. Particularmente, en el caso de que la sonda del vigésimo patrón en la Tabla 2 se seleccione como la sonda de referencia, presenta el valor de coincidencia de 1 (es decir, $a_{20,j}=1$) en todas las posiciones de la sonda. Sin embargo, debido a que la sonda del patrón 17º es no coincidente en las posiciones 18 y 20, $a_{17,18}$ y $a_{17,20}$ fueron representados por 0, y las otras posiciones se representaron con un 1 ($a_{17,j}=1$), debido a que eran coincidentes. La sonda del 18º patrón con una no coincidencia en la 18ª posición ($a_{18,18}=0$) se describió mediante la utilización de una constante binaria para la posibilidad de que fuese coincidente o no coincidente con la sonda del vigésimo patrón (la sonda de referencia). Se encontró que la sonda del 19º patrón era coincidente o no coincidente con la sonda del vigésimo patrón (la sonda de referencia), describiéndose como 0 solo en la posición 19º ($a_{19,19}=0$) y 1 en las demás posiciones. En resumen, a_{ij} de la sonda del vigésimo patrón se expresa como 1 debido a que es perfectamente coincidente con la sonda de referencia, pero a_{ij} de las sondas de los patrones 17º a 19º se expresa como 1 y 0 debido a que incluyen posiciones coincidentes y no coincidentes con la sonda de referencia.

20 En la Tabla 2, en el caso de que se introduzca un número limitado (p. ej., dos) de bases degeneradas en la sonda de referencia para convertirse en perfectamente coincidente con las sondas de los patrones 17º y 18º, se introdujeron bases degeneradas en las posiciones 18 y 20 de la sonda de referencia que debía coincidir con las sondas de los patrones 17º y 18º. La posibilidad de que la sonda de referencia con o sin la base degenerada introducida fuese coincidente o no coincidente en la j -ésima posición con una sonda perteneciente al patrón de la i -ésima secuencia de sondas se representó como una variable binaria: $a_{ij} + d_j$.

Debido a que las sondas de los patrones 17º y 18º eran todos coincidentes con las sondas de referencia tras la introducción de bases degeneradas en las posiciones 18 y 20 de la sonda de referencia, $a_{ij} + d_j$ de las sondas de los patrones 17º y 18º eran todos 1 o más. Por otra parte, debido a que la sonda del patrón 19º todavía no era coincidente en la posición 19, se encontró que $a_{19,19} + d_{19}$ presentaba un valor de cero.

35 Por lo tanto, con el fin de seleccionar los patrones de sonda coincidentes con una sonda de referencia mediante la introducción de bases degeneradas en la sonda de referencia, es necesario que presente un valor mínimo de 1 para $a_{ij} + d_j$ en todas las posiciones. En el caso de que el valor mínimo sea 0, ello implica que los patrones de la sonda son no coincidentes con la sonda de referencia en por lo menos una posición de los patrones de sonda. Tales patrones de sonda no pueden seleccionarse.

40 En otras palabras, una fórmula de restricción para la selección del patrón de la i -ésima sonda en términos de coincidencia/no coincidencia y la introducción de bases degeneradas es $x_i \leq \min \{a_{ij} + d_j, \text{ para todo } j\}$, lo que implica que $x_i \leq a_{ij} + d_j$ debe cumplirse para todo j en la obtención de soluciones mediante la utilización del algoritmo de programación lineal.

45 En donde se aplica la fórmula de restricción a todos los patrones de secuencia en la Tabla 1, cada patrón de secuencia debe cumplir lo siguiente: $x_i \leq a_{i,1} + d_1$, $x_i \leq a_{i,2} + d_2$, $x_i \leq a_{i,3} + d_3$, $x_i \leq a_{i,4} + d_4$, $x_i \leq a_{i,5} + d_5$, $x_i \leq a_{i,6} + d_6$, $x_i \leq a_{i,7} + d_7$, $x_i \leq a_{i,8} + d_8$, $x_i \leq a_{i,9} + d_9$, $x_i \leq a_{i,10} + d_{10}$, $x_i \leq a_{i,11} + d_{11}$, $x_i \leq a_{i,12} + d_{12}$, $x_i \leq a_{i,13} + d_{13}$, $x_i \leq a_{i,14} + d_{14}$, $x_i \leq a_{i,15} + d_{15}$, $x_i \leq a_{i,16} + d_{16}$, $x_i \leq a_{i,17} + d_{17}$, $x_i \leq a_{i,18} + d_{18}$, $x_i \leq a_{i,19} + d_{19}$ y $x_i \leq a_{i,20} + d_{20}$.

50 Se obtuvo una solución óptima del problema de programación lineal utilizando el programa MATLAB R2016b (MathWorks) para conseguir la fórmula objetiva junto con el cumplimiento de las fórmulas de restricción. Se calculó el resultado como $d_2=1$, $d_{18}=1$ y $d_{19}=1$ en las posiciones $j=2$, 18 y 19 y $d_j=0$ en las demás posiciones, y calculado como $x_3=1$, $x_{18}=1$, $x_{19}=1$ y $x_{20}=1$ para los patrones de secuencia $i=3$, 18, 19 y 20 y $x_i=0$ para los demás patrones de secuencia.

Se ha encontrado que al introducir las bases degeneradas en las posiciones $j=2$, 18 y 19 de la sonda de referencia, la sonda de referencia introducida era coincidente con los patrones de secuencia de $i=3$, 18, 19 y 20, el número de sondas incluidas en los patrones de secuencia de $i=3$, 18, 19 y 20 es $p_3=3$, $p_{18}=86$, $p_{19}=4$ y $p_{20}=413$, respectivamente, y 506 sondas del total de 531 sondas eran coincidentes con la sonda de referencia (tasa de coincidencia de 95,3 %).

Tras determinar que es necesario introducir las bases degeneradas en las posiciones 2, 18 y 19 de la sonda de referencia, se introdujo la base degenerada S (G o C) en $j=2$, la base degenerada M (A o C) en $j=18$ y la base degenerada K (G o T) en $j=19$. Utilizando las sondas en las que se habían introducido las bases degeneradas, se verificó que la cobertura diana máxima para la pluralidad de secuencias diana de ácido nucleico en la Tabla 1 era de 95,3 %.

Ejemplo 3: verificación de la solución óptima del problema de programación lineal.

Se verificó la solución óptima del problema de programación lineal obtenida en el Ejemplo 2. En primer lugar, en las secuencias de sonda de la Tabla 1, se utilizó la sonda del patrón de la vigésima secuencia a modo de secuencia de sonda. a_{ij} estaba representada por 1 en el caso de que un patrón de secuencia fuese coincidente con la sonda de referencia y representada por 0 en el caso de que un patrón de secuencia fuese no coincidente. Los valores mínimos de a_{ij} en cada patrón de secuencia se muestran en la Tabla 3 y se resumen los resultados.

Tabla 3

i	Secuencia de oligonucleótido																				p_i	$\min(a_{ij})$
1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
2	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	0
4	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0
5	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
6	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0
7	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	2	0
8	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0
9	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	5	0
10	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0
11	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1	1	3	0
12	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	2	0	
13	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	2	0	
14	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	86	0	
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	4	0	
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	413	1	

Además, la solución del Ejemplo 2, " $d_2=1$, $d_{18}=1$ y $d_{19}=1$ en las posiciones $j=2$, 18 y 19, y $d_j=0$ en la demás posiciones" para la introducción de tres o menos bases degeneradas se introdujo en la Tabla 3 para obtener el valor mínimo de $a_{ij} + d_j$ en cada patrón de secuencia. Se resumen los resultados en la Tabla 4.

Tabla 4

d_j		0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0																			
i	± 1	Secuencia de oligonucleótido																			
1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	2	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	2	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	2	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
10	1	2	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	1	2	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	1	2	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	2	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
14	1	2	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
15	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

A continuación, se seleccionaron los patrones de secuencia que cumplieran $x_i \leq \min \{a_{ij} + d_j\}$. Se muestran los resultados en la Tabla 5, a continuación.

Tabla 5

d_j		0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0																			
i	± 1	Secuencia de oligonucleótido																			
1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	2	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
5	1	2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	2	1	0	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
7	1	2	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
8	1	2	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	2	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
10	1	2	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1
11	1	2	1	1	1	1	0	1	1	1	1	1	0	2	1	0					
12	1	2	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	2	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
14	1	2	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
15	1	2	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
16	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Tal como se muestra en la Tabla 5, se verificó que mediante la introducción de las bases degeneradas en la sonda de referencia, los patrones de secuencia que cumplen $x_i \leq \min \{a_{ij} + d_j\}$ son $i=3$, 18 y 19, además del patrón de la vigésima secuencia, lo que concuerda con los resultados obtenidos en el Ejemplo 2.

Ejemplo comparativo 1: introducción de un número limitado (tres) de bases degeneradas en una sonda para obtener la máxima coincidencia entre sondas de manera empírica y manual.

Con el fin de detectar la pluralidad de secuencias diana de ácido nucleico con una máxima cobertura diana de una manera empírica y manual, se introdujeron tres o menos bases degeneradas en sondas.

En particular, en los 20 patrones de secuencia mostrados en la Tabla 1, debido a que el patrón de la vigésima secuencia incluye 413 secuencias de sonda que presentan el número más grande de secuencias idénticas, y a continuación el patrón de la 18ª secuencia incluye 86 secuencias idénticas, se seleccionaron los patrones de las secuencias 20ª y 18ª. Al introducir la base degenerada en la posición 18 ($j=18$), 499 secuencias de sonda se volvieron coincidentes con la sonda de referencia. Además, al introducir la base degenerada en el patrón de la 9ª secuencia, que presenta el tercer número más grande de secuencias en la Tabla 1, 504 secuencias de sonda se volvieron coincidentes con la sonda de referencia. En este caso, es necesario introducir bases degeneradas en las posiciones 7, 8, 10 y 18 ($j=7, 8, 10$ y 18). Sin embargo, los patrones de la 20ª, 18ª y 9ª secuencias no pueden seleccionarse debido a que la condición de restricción de no introducción de más de 3 bases degeneradas no se cumple.

A continuación, se seleccionaron los patrones de la 20ª, 18ª y 15ª secuencias. Se introdujeron bases degeneradas en las posiciones 16 y 18 ($j=16$ y 18) de manera que 413 secuencias incluidas en el patrón de la 20ª secuencia, 86 secuencias incluidas en el patrón de la 18ª secuencia y 1 secuencia incluida en el patrón de la 15ª secuencia (un total de 500 secuencias de sonda) eran coincidentes con la sonda de referencia.

Mediante la repetición del procedimiento manual, pudo determinarse que al introducir bases degeneradas en las posiciones $j=2, 18$ y 19 de la sonda de referencia, la sonda de referencia era coincidente con los patrones de secuencia $i=3, 18, 19$ y 20 (506 sondas de entre 531 sondas). El método manual requirió aproximadamente 56 veces más tiempo que el método que utilizaba el algoritmo de programación lineal (Ejemplo 2) para determinar las posiciones de introducción de bases degeneradas y la selección de los patrones de secuencia.

REIVINDICACIONES

1. Método implementado por ordenador de preparación de un oligonucleótido para la detección de una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima, que comprende:

(a) seleccionar cada una de las secuencias con similitud de secuencia en la misma región de sondeo de una pluralidad de secuencias diana de ácido nucleico como secuencia de sondeo o como secuencia oligonucleotídica para preparar una pluralidad de secuencias de sondeo o una pluralidad de oligonucleótidos,

(b) seleccionar una secuencia de entre la pluralidad de secuencias de sondeo a modo de secuencia de sondeo de referencia o uno de la pluralidad de oligonucleótidos como oligonucleótido de referencia, y

(c) introducir una base degenerada y/o base universal en un número permisible predeterminado en la secuencia de sondeo de referencia o el oligonucleótido de referencia en una o más posiciones no coincidentes entre (i) la secuencia de sondeo de referencia o el oligonucleótido de referencia, y (ii) la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos para conseguir la máxima coincidencia con la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos, y seleccionar como secuencia de un oligonucleótido, la secuencia de sondeo de referencia o una secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una de la base degenerada y/o la base universal, preparando de esta manera un oligonucleótido con una cobertura diana máxima, y

en el que la etapa (c) se lleva a cabo para conseguir la fórmula objetiva 1 a continuación, junto con el cumplimiento de las fórmulas de restricción 1 y 2 siguientes:

Fórmula objetiva 1

$$\text{Máx.: } \sum_{i=1}^r x_i$$

en donde Máx.: representa maximización; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se ha seleccionado o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos, e " r " es un índice de secuencias de sondeo u oligonucleótidos comprendido entre 1 y r ,

Fórmula de restricción 1:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en donde d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si se ha introducido o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; j es un índice de posiciones de una secuencia de sondeo de referencia o un oligonucleótido de referencia comprendido entre 1 y c ; y $D_{lim.}$ es un número limitado de bases degeneradas y/o bases universales introducidas en la secuencia de sondeo de referencia u oligonucleótido de referencia;

Fórmula de restricción 2:

$$x_i \leq \min\{a_{i,j} + d_j\} \text{ para todo } i, j$$

en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) sobre si se ha seleccionado o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos; $a_{i,j}$ es una constante binaria que representa si la i -ésima secuencia de sondeo es coincidente o no coincidente en la posición j -ésima con la secuencia de sondeo de referencia o el i -ésimo oligonucleótido es coincidente o no coincidente en la j -ésima posición con el oligonucleótido de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; $a_{i,j} + d_j$ es una variable binaria que representa si la secuencia de sondeo de referencia con o sin

la base degenerada y/o la base universal introducida es coincidente o no coincidente en la j -ésima posición con la i -ésima secuencia de sondeo o el oligonucleótido de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con el i -ésimo oligonucleótido; $\{a_{ij} + d_j\}$ representa un conjunto que incluye $a_{ij} + d_j$ como elementos; $\min \{a_{ij} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{ij} + d_j\}$ y para todo i, j representa la aplicación a todas las posiciones de todas las secuencias de sondeo u oligonucleótidos.

2. Método según la reivindicación 1, en donde el método comprende, además, entre las etapas (a) y (b), (a-1) agrupar la pluralidad de secuencias de sondeo o la pluralidad de oligonucleótidos según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, en donde la etapa (c) se lleva a cabo mediante la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la secuencia de sondeo de referencia o en el oligonucleótido de referencia en una o más posiciones no coincidentes entre la secuencia de sondeo de referencia o el oligonucleótido de referencia y la pluralidad de patrones de secuencia que van a hacerse coincidir al máximo con la pluralidad de patrones de secuencia, preparando de esta manera el oligonucleótido con la máxima cobertura diana.

3. Método según la reivindicación 1 o 2, en el que la secuencia de sondeo de referencia es una secuencia de sondeo que presenta el número más alto de secuencias idénticas de la pluralidad de secuencias de sondeo, o una secuencia de sondeo de un patrón de secuencia que presenta el número más alto de secuencias de sondeo agrupadas en un patrón de secuencia de la pluralidad de patrones de secuencia, y en el que el oligonucleótido de referencia es un oligonucleótido que presenta el número más alto de secuencias idénticas de la pluralidad de oligonucleótidos, o un oligonucleótido de un patrón de secuencia que presenta el número más alto de oligonucleótidos agrupados en un patrón de secuencia de la pluralidad de patrones de secuencia.

4. Método según la reivindicación 2, en donde la etapa (c) se lleva a cabo para conseguir la fórmula objetiva 2 a continuación, junto con el cumplimiento de las fórmulas de restricción 3 y 4 siguientes:

Fórmula objetiva 2

$$\text{Máx.: } \sum_{i=1}^r x_i$$

en donde Máx.: representa la maximización; p_i es el número de secuencias de sondeo pertenecientes al patrón de la i -ésima secuencia de las secuencias de sondeo o es el número de oligonucleótidos pertenecientes al patrón de la i -ésima secuencia de oligonucleótidos; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no el patrón de la i -ésima secuencia de las secuencias de sondeo o el patrón de la i -ésima secuencia de oligonucleótidos, e " i " es un índice de patrones de secuencia de las secuencias de sondeo u oligonucleótidos comprendido entre 1 y r ,

Fórmula de restricción 3:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en la que d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; j es un índice de posiciones de la secuencia de sondeo de referencia o el oligonucleótido de referencia comprendido entre 1 y c , y D_{lim} es un número limitado de bases degeneradas y/o bases universales introducidas en la secuencia de sondeo de referencia o del oligonucleótido de referencia;

Fórmula de restricción 4:

$$x_i \leq \min\{a_{ij} + d_j\} \text{ para todo } i, j$$

en la que x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para la selección o no del patrón de la i -ésima secuencia de las secuencias de sondeo o el patrón de la i -ésima secuencia de oligonucleótidos; a_{ij} es una constante binaria que representa si una secuencia de sondeo perteneciente al patrón de la i -ésima secuencia de secuencias de sondeo es coincidente o no coincidente en la j -ésima posición con la secuencia de sondeo de referencia o un oligonucleótido

perteneciente al patrón de la i -ésima secuencia de oligonucleótidos es coincidente o no coincidente en la j -ésima posición con el oligonucleótido de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o la base universal en la j -ésima posición de la secuencia de sondeo de referencia o del oligonucleótido de referencia; $a_{i,j} + d_j$ es una variable binaria que representa si la secuencia de sondeo de referencia con o sin la base degenerada introducida y/o base universal es coincidente o no coincidente en la j -ésima posición con una secuencia de sondeo perteneciente al patrón de la i -ésima secuencia de las secuencias de sondeo o el oligonucleótido de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con el oligonucleótido perteneciente al patrón de la i -ésima secuencia de oligonucleótidos; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min \{a_{i,j} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{i,j} + d_j\}$; y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo u oligonucleótidos.

5. Método según la reivindicación 1 o 4, en donde el valor de no selección ($x_{no\ sel.}$) y el valor de selección ($x_{sel.}$) de x_i son 0 y 1, respectivamente; en donde el valor de no introducción ($d_{no\ int.}$) y el valor de introducción ($d_{int.}$) de d_j son 0 y 1, respectivamente; en donde D_{lim} es 3, o en donde el valor de no coincidencia y el valor de coincidencia de $a_{i,j}$ son 0 y 1, respectivamente.

6. Método según la reivindicación 1, en donde la pluralidad de secuencias diana de ácido nucleico son una pluralidad de secuencias de ácido nucleico que presentan similitud de secuencia para una molécula diana de ácido nucleico que muestra diversidad genética, o en donde la pluralidad de secuencias diana de ácido nucleico son una pluralidad de secuencias de ácido nucleico correspondientes a homólogos de organismos que presentan la misma función, la misma estructura o el mismo nombre génico.

7. Método implementado por ordenador para determinar una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia, que comprende:

(a) seleccionar una región de sondeo candidata en una pluralidad de secuencias diana de ácido nucleico y proporcionar secuencias en la región de sondeo candidata en forma de una pluralidad de secuencias de sondeo,

(b) introducir una base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre la pluralidad de secuencias de sondeo, de manera que la pluralidad de secuencias de sondeo coinciden al máximo entre sí, y

(c) determinar la región de sondeo candidata como una región de sondeo para detectar la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico en el caso de que la coincidencia máxima sea no inferior a una cobertura predeterminada, y

en donde la etapa (b) se lleva a cabo para conseguir la fórmula objetiva 3 a continuación, junto con el cumplimiento de las fórmulas de restricción 5 y 6 siguientes:

Fórmula objetiva 3

$$\text{Máx.:} \sum_{i=1}^r x_i$$

en donde Máx.: representa maximización; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo, e " r " es un índice de secuencias de sondeo comprendido entre 1 y r ,

Fórmula de restricción 5

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en la que d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la pluralidad de secuencias de sondeo; j es un índice de posiciones de la secuencia de sondeo de referencia comprendido entre 1 y c ; y D_{lim} es un número limitado de la base degenerada y/o base universal introducida en la pluralidad de secuencias de sondeo;

Fórmula de restricción 6

$$x_i \leq \min\{a_{i,j} + d_j\} \text{ para todo } i, j$$

en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo; $a_{i,j}$ es una constante binaria que representa si la pluralidad de secuencias de sondeo son coincidentes o no coincidentes entre sí en la posición j -ésima; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la pluralidad de secuencias de sondeo; $a_{i,j} + d_j$ es una variable binaria que representa si la pluralidad de secuencias de consenso con o sin la base degenerada y/o base universal introducida son coincidentes o no coincidentes entre sí en la j -ésima posición; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min\{a_{i,j} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{i,j} + d_j\}$ y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo.

8. Método según la reivindicación 7, en donde la etapa (b) se lleva a cabo mediante la selección de una de la pluralidad de secuencias de sondeo a modo de secuencia de sondeo de referencia, y la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la secuencia de sondeo de referencia en una o más posiciones no coincidentes entre la secuencia de sondeo de referencia y la pluralidad de secuencias de sondeo para conseguir la máxima coincidencia con la pluralidad de secuencias de sondeo.

9. Método según la reivindicación 7, en donde el método comprende, además, entre las etapas (a) y (b), (a-1) la agrupación de la pluralidad de secuencias de sondeo según identidad de secuencia a fin de obtener una pluralidad de patrones de secuencia, en donde la etapa (b) se lleva a cabo mediante la introducción de la base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre sí en la pluralidad de patrones de secuencia de manera que la pluralidad de patrones de secuencia son coincidentes al máximo entre sí, y en donde la etapa (c) se lleva a cabo mediante la determinación de la región de sondeo candidata a la que pertenece el patrón de secuencia a modo de región de sondeo para detectar la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico en el caso de que la coincidencia máxima sea no inferior a la cobertura predeterminada.

10. Método según la reivindicación 8, en donde la secuencia de sondeo de referencia es una secuencia de sondeo que presenta el número más alto de secuencias idénticas de la pluralidad de secuencias de sondeo o una secuencia de sondeo de un patrón de secuencia que presenta el número más alto de secuencias de sondeo de la pluralidad de patrones de secuencia.

11. Método según la reivindicación 9, en donde, en la fórmula de restricción 5, d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia; j es un índice de posiciones de la secuencia de sondeo de referencia comprendido entre 1 y c ; y D_{lim} es un número limitado de la base degenerada y/o base universal introducida en la secuencia de sondeo de referencia, y en donde, en la fórmula de restricción 6, $a_{i,j}$ es una constante binaria que representa si la i -ésima secuencia de sondeo es coincidente o no coincidente en la j -ésima posición con la secuencia de sondeo de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia; y $a_{i,j} + d_j$ es una variable binaria que representa si la secuencia de sondeo de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con la i -ésima secuencia de sondeo.

12. Método según la reivindicación 9, en donde la etapa (b) se lleva a cabo para conseguir la fórmula objetiva 4 a continuación, junto con el cumplimiento de las fórmulas de restricción 7 y 8 siguientes:

Fórmula objetiva 4

$$\text{Máx. : } \sum_{i=1}^r p_i x_i$$

en donde Max: representa maximización; p_i es el número de secuencias de sondeo pertenecientes al patrón de la i -ésima secuencia de secuencias de sondeo; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no el patrón de la i -ésima secuencia de las secuencias de sondeo, e " r " es un índice de patrones de secuencia de las secuencias de sondeo

comprendido entre 1 y r ,

Fórmula de restricción 7:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en donde d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo del patrón de secuencia o la secuencia de sondeo de referencia; j es un índice de posiciones de la secuencia de sondeo del patrón de secuencia o la secuencia de sondeo de referencia comprendido entre 1 y c ; y D_{lim} es un número limitado de la base degenerada y/o base universal introducida en la secuencia de sondeo del patrón de secuencia o la secuencia de sondeo de referencia;

Fórmula de restricción 8:

$$x_i \leq \min\{a_{i,j} + d_j\} \text{ para todo } i, j$$

en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no el patrón de la i -ésima secuencia de secuencias de sondeo; $a_{i,j}$ es una constante binaria que representa si las secuencias de sondeo de cada uno de los patrones de secuencia son coincidentes o no coincidentes entre sí en la j -ésima posición o una secuencia de sondeo del patrón de la i -ésima secuencia es coincidente o no coincidente en la j -ésima posición con la secuencia de sondeo de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de cada uno de los patrones de secuencia o la secuencia de sondeo de referencia; $a_{i,j} + d_j$ es una variable binaria que representa si las secuencias de sondeo de cada uno de los patrones de secuencia con o sin la base degenerada y/o base universal introducida son coincidentes o no coincidentes entre sí en la j -ésima posición o la secuencia de sondeo de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con la secuencia de sondeo del patrón de la i -ésima secuencia; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min\{a_{i,j} + d_j\}$ representa un valor mínimo de los elementos del conjunto $\{a_{i,j} + d_j\}$ y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo.

13. Medio de almacenamiento legible por ordenador que contiene instrucciones para configurar un procesador a fin de que lleve a cabo un método de preparación de un oligonucleótido para la detección de una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia con una cobertura diana máxima, en donde el método comprende:

- (a) seleccionar cada una de las secuencias con similitud de secuencia en la misma región de una pluralidad de secuencias diana de ácido nucleico a modo de oligonucleótido para la preparación de una pluralidad de oligonucleótidos,
- (b) seleccionar uno de la pluralidad de oligonucleótidos a modo de oligonucleótido de referencia, y
- (c) introducir una base degenerada y/o base universal en un número permisible predeterminado en el oligonucleótido de referencia en una o más posiciones de no coincidencia entre el oligonucleótido de referencia y la pluralidad de oligonucleótidos que van a hacerse coincidir al máximo con la pluralidad de oligonucleótidos, y seleccionar a modo de secuencia de un oligonucleótido, una secuencia del oligonucleótido de referencia en la que se ha introducido por lo menos una de la base degenerada y/o la base universal, preparando de esta manera un oligonucleótido con una cobertura diana máxima, y

en donde la etapa (c) se lleva a cabo para conseguir la fórmula objetiva 1 a continuación, junto con el cumplimiento de las fórmulas de restricción 1 y 2 siguientes:

Fórmula objetiva 1:

$$\text{Máx.: } \sum_{i=1}^r x_i$$

en donde Máx.: representa maximización; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) de si selecciona o no la i -ésima secuencia de sondeo de la pluralidad de

secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos, e " r " es un índice de secuencias de sondeo u oligonucleótidos comprendido entre 1 y r ,

Fórmula de restricción 1:

$$\sum_{j=1}^c d_j \leq D_{lim}$$

en donde d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si se introduce o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; j es un índice de posiciones de una secuencia de sondeo de referencia o un oligonucleótido de referencia comprendido entre 1 y c ; y $D_{lim.}$ es un número limitado de la base degenerada y/o base universal introducida en la secuencia de sondeo de referencia u oligonucleótido de referencia;

Fórmula de restricción 2:

$$x_i \leq \min\{a_{i,j} + d_j\} \text{ para todo } i, j$$

en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo o el i -ésimo oligonucleótido de la pluralidad de oligonucleótidos; $a_{i,j}$ es una constante binaria que representa si la i -ésima secuencia de sondeo es coincidente o no coincidente en la posición j -ésima con la secuencia de sondeo de referencia o el i -ésimo oligonucleótido es coincidente o no coincidente en la j -ésima posición con el oligonucleótido de referencia; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la secuencia de sondeo de referencia o el oligonucleótido de referencia; $a_{i,j} + d_j$ es una variable binaria que representa si la secuencia de sondeo de referencia con o sin la base degenerada y/o la base universal introducida es coincidente o no coincidente en la j -ésima posición con la i -ésima secuencia de sondeo o el oligonucleótido de referencia con o sin la base degenerada y/o base universal introducida es coincidente o no coincidente en la j -ésima posición con el i -ésimo oligonucleótido; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min\{a_{i,j} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{i,j} + d_j\}$ y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo u oligonucleótidos.

14. Medio de almacenamiento legible por ordenador que contiene instrucciones para configurar un procesador para llevar a cabo un método de determinación de una región de sondeo en una pluralidad de secuencias diana de ácido nucleico que presentan similitud de secuencia, en donde el método comprende:

- (a) seleccionar una región de sondeo candidata en una pluralidad de secuencias diana de ácido nucleico y proporcionar secuencias en la región de sondeo candidata en forma de una pluralidad de secuencias de sondeo,
- (b) introducir una base degenerada y/o base universal en un número permisible predeterminado en la pluralidad de secuencias de sondeo en una o más posiciones no coincidentes entre la pluralidad de secuencias de sondeo, de manera que la pluralidad de secuencias de sondeo sean coincidentes al máximo entre sí, y
- (c) determinar la región de sondeo candidata a modo de región de sondeo para la detección de la totalidad o una parte de la pluralidad de secuencias diana de ácido nucleico en el caso de que la coincidencia máxima sea no inferior a una cobertura predeterminada, y

en donde la etapa (b) se lleva a cabo para conseguir la fórmula objetiva 3 a continuación, junto con el cumplimiento de las fórmulas de restricción 5 y 6 siguientes:

Fórmula objetiva 3

$$\text{Máx.: } \sum_{i=1}^r x_i$$

en donde Máx.: representa maximización; x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo, e " r " es un índice de secuencias de sondeo comprendidas entre 1 y r ,

Fórmula de restricción 5

$$\sum_{j=1}^c d_j \leq D_{lim}$$

5 en la que d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la pluralidad de secuencias de sondeo; j es un índice de posiciones de la secuencia de sondeo de referencia comprendido entre 1 y c ; y D_{lim} es un número limitado de la base degenerada y/o base universal introducida en la pluralidad de secuencias de sondeo;

Fórmula de restricción 6

$$x_i \leq \min\{a_{i,j} + d_j\} \text{ para todo } i, j$$

15 en donde x_i es una variable binaria que consiste en un valor de no selección ($x_{no\ sel.}$) y un valor de selección ($x_{sel.}$) para si se selecciona o no la i -ésima secuencia de sondeo de la pluralidad de secuencias de sondeo; $a_{i,j}$ es una constante binaria que representa si la pluralidad de secuencias de sondeo son coincidentes o no coincidentes entre sí en la posición j -ésima; d_j es una variable binaria que consiste en un valor de no introducción ($d_{no\ int.}$) y un valor de introducción ($d_{int.}$) que representa si introducir o no la base degenerada y/o base universal en la j -ésima posición de la pluralidad de secuencias de sondeo; $a_{i,j} + d_j$ es una variable binaria que representa si la pluralidad de secuencias de consenso con o sin la base degenerada y/o base universal introducida son coincidentes o no coincidentes entre sí en la j -ésima posición; $\{a_{i,j} + d_j\}$ representa un conjunto que incluye $a_{i,j} + d_j$ como elementos; $\min\{a_{i,j} + d_j\}$ representa un valor mínimo entre los elementos del conjunto $\{a_{i,j} + d_j\}$ y para todo i, j representa la aplicación de todas las posiciones de todas las secuencias de sondeo.

Fig. 1

$$A = \{a_{ij}\} =$$

Fig. 2

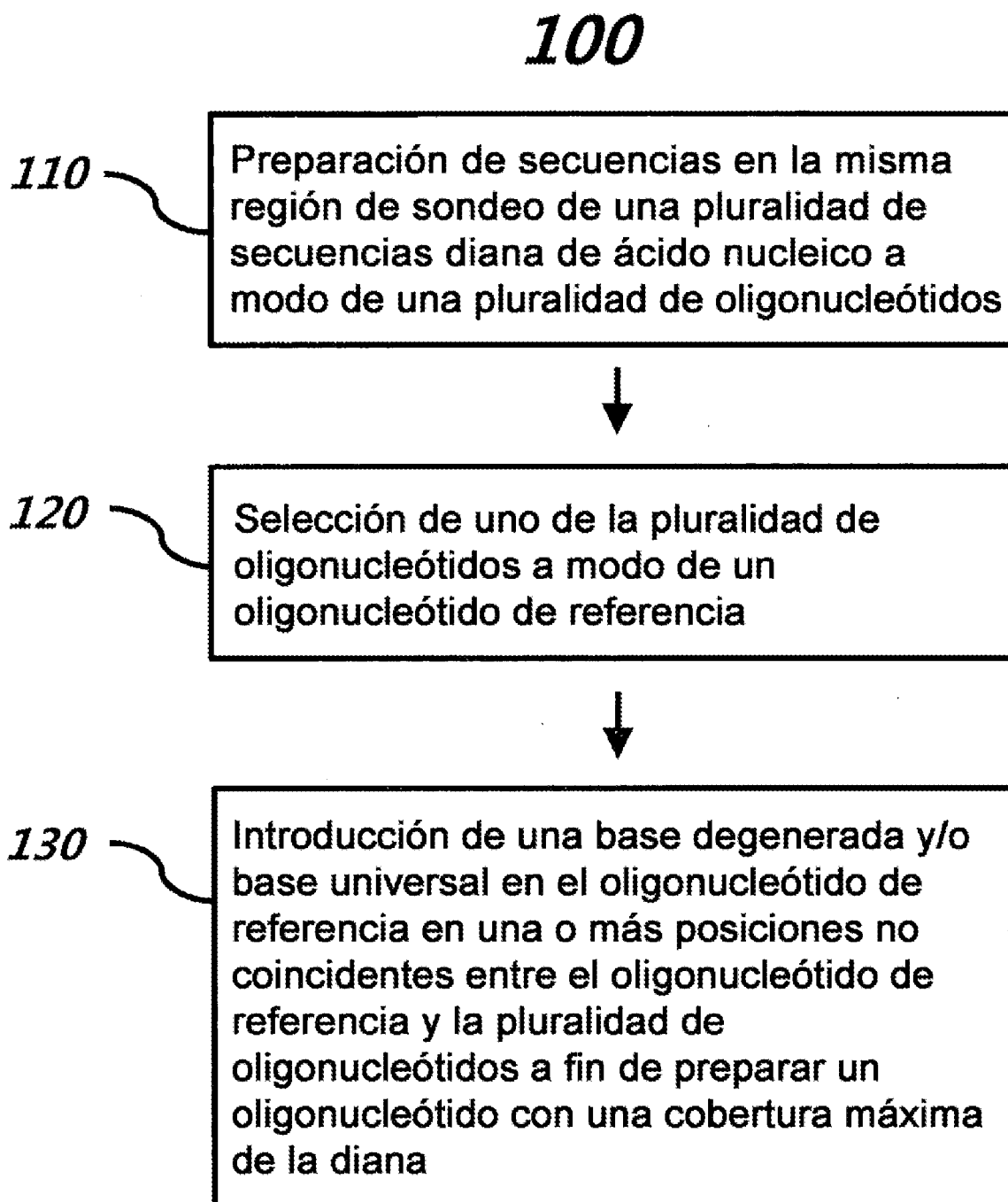


Fig. 3

