



19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA

11 Número de publicación: **2 281 426**

51 Int. Cl.:
G06F 19/00 (2006.01)
G06F 17/00 (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Número de solicitud europea: **01949637 .1**
86 Fecha de presentación : **02.07.2001**
87 Número de publicación de la solicitud: **1297484**
87 Fecha de publicación de la solicitud: **02.04.2003**

54 Título: **Procedimiento para el análisis de secuencias.**

30 Prioridad: **05.07.2000 GB 0016472**

45 Fecha de publicación de la mención BOPI:
01.10.2007

45 Fecha de la publicación del folleto de la patente:
01.10.2007

73 Titular/es: **GE Healthcare UK Limited**
Amersham Place
Little Chalfont, Buckinghamshire HP7 9NA, GB

72 Inventor/es: **Odedra, Raj**

74 Agente: **Carpintero López, Francisco**

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

Procedimiento para el análisis de secuencias.

5 **Campo de la invención**

La presente invención se refiere a un procedimiento de secuenciación y aparato que permite la corrección de errores durante la secuenciación de moléculas individuales.

10 **Antecedentes de la invención**

La secuenciación se realiza de manera rutinaria mediante el procedimiento de la terminación de cadenas y separación en gel, esencialmente tal como se describe por Sanger, F., S. Nicklen, y A. Coulson (Proc Natl Acad Sci USA, 1977, 74(12); págs. 5463-7). El procedimiento se basa en la generación de una población mixta de fragmentos de ADN que representan terminaciones en cada base en la secuencia. Entonces se determina la secuencia mediante separación electroforética de estos fragmentos.

Esfuerzos recientes para aumentar el rendimiento de secuenciación han dado como resultado el desarrollo de procedimientos alternativos que eliminan la etapa de separación electroforética. Varios de estos procedimientos utilizan la extensión de bases (es decir, adición de bases) y se han descrito, por ejemplo, en los documentos WO 93/21340, US 5.302.509 y US 5.547.839. En estos procedimientos, los moldes o cebadores se inmovilizan sobre una superficie sólida antes de la exposición a reactivos para su secuenciación. Las moléculas inmovilizadas se incuban en presencia de análogos de nucleótidos que tienen una modificación en el carbono en 3' del residuo de azúcar que bloquea reversiblemente el grupo hidroxilo en esa posición. La incorporación de tales nucleótidos modificados por una polimerasa garantiza que sólo se añade un nucleótido durante cada ciclo de extensión de bases. Entonces se detecta la base añadida en virtud de una etiqueta que se ha incorporado en el grupo bloqueante en 3'. Tras la detección, el grupo bloqueante se elimina (o "corta"), normalmente por medios fotoquímicos para exponer un grupo hidroxilo libre que está disponible para la adición de bases durante el siguiente ciclo.

Generalmente, los enfoques no basados en la separación se basan en la presencia de grandes números de moléculas de molde para cada secuencia diana para generar una secuencia consenso a partir de una diana dada. Por tanto, por ejemplo, pueden aplicarse reacciones de extensión de bases a múltiples moldes analizando puntos diferenciados de ácido nucleico, que comprenden cada uno una multiplicidad de moléculas, inmovilizado en una red espacialmente direccionable.

Sin embargo, las reacciones de incorporación de terminador/corte, o escisión de bases son propensas a errores. Por ejemplo, tal como se describió anteriormente, las estrategias de extensión de bases han utilizado generalmente análogos de nucleótidos que combinan las funciones de una molécula indicadora, normalmente un flúor, con las de un terminador que ocupa la posición en 3' en un resto de azúcar. La naturaleza voluminosa del grupo y su posición hacen estos compuestos sustratos sumamente ineficaces para polimerasas. Además, el corte del grupo terminador para permitir las adiciones posteriores también está sometido a ineficacias. En presencia de miles, o preferiblemente millones, de moléculas para cada diana, incluso errores modestos de menos del 5% dan como resultado una pérdida acumulativa de sincronía, entre la multiplicidad de las cadenas que representan cada molécula, dentro de un pequeño número de ciclos. Por tanto, con cada ciclo de secuenciación el ruido de fondo aumenta progresivamente con un deterioro consecuente de la señal con cada adición. Esto significa que el número de bases de datos de secuencia que pueden obtenerse está limitado antes de que la señal específica no pueda distinguirse del fondo.

Avances recientes en procedimientos de detección de moléculas individuales (descritos, por ejemplo, en Trabesinger, W., *et al.*, Anal Chem., 1999, 71(1); págs. 279-83 y el documento WO 00/06770) hacen posible aplicar estrategias de secuenciación a moléculas individuales. Sin embargo, la secuenciación, cuando se aplica a poblaciones clonales de moléculas, es un procedimiento estocástico que da como resultado que algunas moléculas experimenten reacciones mientras que otras permanezcan sin modificarse. Por tanto, en procedimientos de secuenciación convencionales, los errores tales como incorporaciones erróneas no son normalmente de significación grave ya que los grandes números de moléculas presentes garantizan que se obtiene una señal consenso. Cuando estas reacciones se aplican a moléculas individuales, los resultados se cuantifican eficazmente.

Uno de tales procedimientos de secuenciación de moléculas individuales se basa en la escisión de bases y se describe, por ejemplo, en Hawkins, G. y L. Hoffman, Nature Biotechnology, 1997, vol. 15; págs. 803-804 y el documento US 5.674.743. Con esta estrategia, se generan moléculas de molde individuales de tal manera que cada base está marcada con un indicador apropiado. Las moléculas de molde se digieren con exonucleasa y se monitorizan e identifican las bases escindidas. Ya que estos procedimientos utilizan enzimas sumamente procesivas tales como Lambda exonucleasa, hay potencial para analizar moldes grandes de varios kilobases de longitud. Sin embargo, la monitorización continua de bases escindidas a partir de cada molécula de molde en tiempo real limita el número de moléculas que pueden analizarse en paralelo. Además, hay dificultades para generar un molde en el que cada base está marcada con un indicador apropiado de tal manera que las bases escindidas pueden detectarse basándose en las propiedades ópticas o químicas intrínsecas.

También se han adaptado procedimientos basados en la extensión de bases (tales como BASS) al enfoque de moléculas únicas.

Sin embargo, estas técnicas son propensas a errores. En particular, la incorporación de nucleótidos modificados puede fallar, por ejemplo, como resultado de la disminución de la eficacia de la acción de la polimerasa con nucleótidos modificados. Cuando la molécula indicadora es una molécula fluorescente, también pueden producirse errores mediante el fallo de la fluorescencia debido a que se pierde, se daña, se blanquea o se desexcita el flúor. Al nivel de moléculas únicas, los fallos tales como estos darán como resultado que no pueda obtenerse una secuencia adecuada.

Es un objeto de la presente invención proporcionar un procedimiento de secuenciación que permite que se detecten los errores. Es un objeto adicional de la presente invención permitir el análisis y la prevención de errores, o corrección, monitorizando el destino de moléculas individuales mediante reacciones de secuenciación.

Sumario de la invención

La invención en sus diversos aspectos se define en las reivindicaciones independientes siguientes, a las que debe hacerse referencia ahora. Se exponen características ventajosas en las reivindicaciones adjuntas.

Brevemente, en una realización preferida de la invención que toma la forma de un procedimiento de análisis de una secuencia de nucleótidos, se obtiene una secuencia de bases a partir de un molde, y se identifica una base en la secuencia como una base desconocida. Se incluye un indicador "desconocido" en la secuencia en la posición correspondiente a la base desconocida, y se genera una secuencia de salida que contiene el indicador de base desconocido. En la realización preferida, la secuencia de bases se obtiene a partir del molde mediante evaluación de un indicador y asignación de las bases según el mismo. Se realiza una determinación sobre si el indicador procede de un ciclo precedente de determinación de bases, y si el indicador procede de un ciclo precedente de determinación de bases, se descarta la asignación de bases.

La secuencia de nucleótidos que debe analizarse puede ser una secuencia de ARN o de ADN.

Breve descripción del dibujo

Ahora se describirá la invención en más detalle a modo de ejemplo con referencia al dibujo adjunto en el que:

la figura 1 es un diagrama de flujo que ilustra un procedimiento de análisis de datos obtenidos durante una reacción para determinar la secuencia de una molécula biológica, tal como una molécula de ácido nucleico, y formar una realización preferida de la invención.

Descripción detallada de la realización preferida

La figura 1 muestra un diagrama que ejemplifica un procedimiento de obtención de una información de secuencia a partir de un molde. El procedimiento compensa los errores (a) identificando bases que se transfieren a partir de un ciclo anterior y (b) detectando moléculas detenidas que pueden producirse a partir de un fallo de marcaje o incorporación errónea de bases. El procedimiento de análisis de datos emplea una reacción de secuenciación convencional que se realiza tal como sigue. En primer lugar, una molécula de ácido nucleico para la que se requieren datos de secuencia, un molde, se une a una superficie sólida tal como un portaobjetos de microscopio. El molde puede marcarse de manera que pueda determinarse su posición cuando se observa el portaobjetos a través de un explorador de microscopio fluorescente, por ejemplo. Se consulta la primera base o nucleótido, es decir, A, C, G, o T, en la secuencia del molde mediante reacción química añadiendo una base fluorescentemente marcada o una etiqueta que representa esa base. Puede ser una de A, C, G o T, o las cuatro marcadas con cuatro etiquetas distinguibles diferentes. La primera base en el molde se unirá a su base complementaria de una manera bien conocida; es decir, A se une a T, y C se une a G, y viceversa. La incorporación de bases puede realizarse extendiendo el molde con una enzima de polimerasa o ligando un oligonucleótido marcado con una ligasa. La incorporación de la base marcada se detecta y se determina su identidad. Entonces se elimina la etiqueta de esa base. Entonces se repite esta serie de etapas para las bases sucesivas en el molde.

Reacciones de secuenciación convencionales adecuadas que implican incorporación/adición de bases incluyen reacciones de extensión de bases tales como las descritas en los documentos WO 93/21340, US 5.302.509 y US 5.547.839 y técnicas tales como las descritas en los documentos US 5.763.175, US 5.599.675, US 5.856.093 y US 5.715.330 en los que las rondas de secuenciación sucesivas implican escisión de bases del molde antes de la incorporación de la base posterior.

Cuando se realiza esta reacción de secuenciación, pueden producirse errores. Por ejemplo, (i) puede incorporarse incorrectamente una base, es decir, se incorpora erróneamente, o (ii) puede no conseguir eliminarse una etiqueta de un ciclo antes de realizar el siguiente ciclo, o (iii) puede fallar la incorporación de una base en un ciclo cualquiera. En la realización preferida de la invención que va a describirse, los datos de las reacciones de secuencia se asimilan de tal manera que los efectos de estos errores pueden reducirse.

Los procedimientos para la deposición y fijación de moléculas sobre fases sólidas se conocen bien en la técnica. Los procedimientos para unir ácidos nucleicos, por ejemplo, se revisan en Schena (ed.), DNA Microarrays: A practical

approach, Oxford University Press (1999) ISBN: 0199637768. Normalmente, la fase sólida será vidrio, aunque pueden usarse otros materiales tales como plásticos o silicio amorfo o cristalino.

Puede unirse una pluralidad de moléculas a la fase sólida en una red ordenada pero, más preferiblemente, se unirán de una manera aleatoria. Una unión aleatoria de moléculas puede comprender cualquier número de moléculas, distribuidas preferiblemente con una densidad apropiada para la resolución óptica de la información de secuencia.

Un resto indicador adecuado puede ser uno cualquiera de varios sistemas indicadores conocidos. Puede ser un radioisótopo por medio del que se hace fácilmente detectable el análogo de nucleósido incorporado, por ejemplo ^{32}P , ^{33}P , ^{35}S incorporados en un grupo fosfato o tiofosfato o H-fosfonato o alternativamente ^3H o ^{14}C o un isótopo de yodo. Puede ser un isótopo detectable mediante espectrometría de masas o RMN. Puede ser un resto de señal, por ejemplo, una enzima, hapteno, fluoróforo, cromóforo, grupo quimioluminiscente, etiqueta de Raman, etiqueta electroquímica, o compuesto señal adaptado para la detección mediante espectrometría de masas.

Cada etapa de secuenciación dará como resultado la unión de moléculas indicadoras a moldes individuales y la detección del resto indicador incorporado permitirá asignar la identidad de la base. En el caso de indicadores fluorescentes, estas moléculas se identificarán entonces, por ejemplo, mediante microscopía de fluorescencia (por ejemplo, usando PMT o CCD) y la propiedad de fluorescencia del indicador permitirá la asignación de identidad a la base incorporada en la reacción de secuenciación.

Con el fin de recoger datos a partir de rondas secuenciales de ciclos de secuenciación, debe localizarse el molde. Esto puede lograrse simultáneamente con el primer ciclo de secuenciación en el que la molécula indicadora en la primera base identifica la ubicación del molde o el propio molde y/o cebador puede estar marcado con un resto indicador de tal manera que puede detectarse su ubicación sobre la fase sólida mucho antes que la reacción de ciclación de la secuencia. Conocer la ubicación de cada molécula de molde hace posible monitorizar el estado de cada molécula tras todos los acontecimientos posteriores durante los ciclos de secuenciación. El fallo de adición posterior, por ejemplo, se manifiesta por sí mismo mediante una falta de fluorescencia en una ubicación que se sabe que contiene un molde. El fallo del indicador debido o bien a una falta de estímulo, o bien a un daño químico, también puede determinarse una vez que se ha determinado la ubicación del molde. Estas reacciones fallidas pueden rastrearse y tratarse en la secuencia final como posibles huecos debido a fallo del indicador. Si estas moléculas reanudan la participación en ciclos posteriores, esto también puede rastrearse y obtenerse una secuencia con sentido. Pueden identificarse puntos individuales de huecos de una única base y, cuando se han dispuesto en red múltiples secuencias idénticas sobre la superficie sólida, puede construirse una secuencia consenso mediante comparaciones con cadenas de referencia tales como secuencias de otras copias de moldes en la red de secuenciación. Alternativamente, pueden identificarse huecos de una única base mediante comparación con una cadena de referencia que puede ser la secuencia conocida (por ejemplo en la aplicación de esta técnica a la detección de mutaciones).

Por tanto, se ha apreciado que en este sistema es posible corregir errores, particularmente errores asociados a la secuenciación de moléculas individuales. Los errores que necesitan corregirse son el fallo en el corte y eliminación de indicador antes del siguiente ciclo, fallo de incorporación, daño al indicador (por ejemplo, daño al flúor), e incorporación errónea.

Una vez localizada, todos los resultados de ciclo de secuenciación para la molécula localizada podrán medirse. El uso de dos conjuntos de análogos de nucleótidos permite la identificación del indicador que se ha transferido desde un ciclo anterior. Por tanto, la recurrencia de un indicador del ciclo anterior puede identificarse y monitorizarse.

Conocer la ubicación de la molécula de molde también permite la identificación de moldes que no parecen haberse extendido. Tal como se trató anteriormente, el fallo para observar una molécula de indicador puede deberse a la falta de incorporación, pero también puede deberse al daño al resto indicador. Sin embargo, ya que puede minimizarse eficazmente la presencia de moléculas dañadas mediante un procedimiento de purificación durante la síntesis de nucleótidos modificados en los que pueden identificarse y eliminarse productos de ruptura y productos de reacciones secundarias, la ausencia de fluorescencia se debe por tanto más posiblemente a un resultado de fallo al incorporar un nucleótido modificado.

Si, tras cualquier ciclo de secuenciación, una molécula de molde no está asociada con ningún indicador, la secuencia se marca en consecuencia en este punto para indicar una "detención". En la siguiente ronda de secuenciación, la molécula de molde puede asociarse entonces con un indicador, es decir, la molécula "detenida" reanuda la extensión permitiendo obtenerse datos de secuencia. Sin embargo, la molécula de molde puede seguir careciendo de asociación con ningún indicador durante más de un ciclo, y la secuencia se marcará como una detención para cada ciclo respectivo.

Un marcador posicional generado durante la secuenciación será útil para interpretar huecos en las alineaciones cuando se compare con la secuencia generada con secuencias de referencia o con otras secuencias generadas durante el procedimiento de secuenciación usando uno de los algoritmos de alineación conocidos por los expertos en la técnica.

Es posible predecir posiciones de incorporación errónea conociendo las propiedades inherentes de las ligasas y polimerasas pertinentes usadas. Por ejemplo, los expertos en la técnica conocen que las secuencias cebadoras que contienen una base terminal con apareamiento erróneo son moldes peores para polimerasas, con eficacias de extensión

de entre 10^2 a 10^6 veces inferiores a las secuencias apareadas (véase Huang, M., N. Arnheim, y M. Goodman, *Nucleic Acids Res*, 1992. 20(17): págs. 4567-73., Tindall KR, K. T., *Biochemistry*, 1988. 27(16): págs. 6008-13, Esteban, J., M. Salas, y L. Blanco, *J Biol Chem*, 1993. 268(4): págs. 2719-26). Las moléculas que permanecen detenidas durante varios ciclos, o al final del protocolo de secuenciación, tienen por tanto una posibilidad mucho mayor de contener un apareamiento erróneo terminal. Por tanto, se marcan los moldes que experimentan tales detenciones en la posición de llamada de última base como terminaciones posibles debidas a apareamientos erróneos. La identificación del fragmento secuenciado se logra entonces mediante alineamiento con una secuencia de referencia y otros moldes secuenciados a partir de la misma muestra. Los apareamientos erróneos que se producen en posiciones marcadas es más probable que resulten de incorporaciones erróneas en vez de representar la verdadera secuencia y por tanto pueden interpretarse en consecuencia.

El número de ciclos para el que se detiene una molécula de molde puede contarse mediante detección sucesiva de una falta de indicador incorporado. Puede fijarse un umbral para la posibilidad de detenciones sucesivas resultantes por casualidad durante en análisis de los datos de secuencia. El umbral sobre el que las detenciones sucesivas pueden clasificarse como que resultan de un apareamiento erróneo dependerá de la eficacia de marcaje o bien mediante extensión de bases dependiente de polimerasa, o bien mediante ligación dependiente de secuencia. Por ejemplo, si el umbral para la posibilidad de detenciones sucesivas resultantes por casualidad se fija al $1 \times 10^{-6}\%$ se contarán los siguientes números de detenciones, teniendo en cuenta diferentes eficacias de marcaje, antes de contar la detención como un apareamiento erróneo.

Eficacia de marcaje por extensión de cebador, o ligación de términos cohesivos	Número de detenciones encontradas antes de superar un punto de corte de posibilidad del $1 \times 10^{-6}\%$
99,9%	3
99,5%	4
99%	4
95%	5
90%	6
80%	8

Para una mayor certeza, puede aumentarse el umbral apropiadamente. El grado de certeza requerido dependerá de la tolerancia de la aplicación de secuenciación; puede tolerarse un punto de corte menos riguroso si el objetivo es simplemente identificar los fragmentos de molde, en vez de determinar con precisión las diferencias de secuencia. El efecto de una eficacia inferior de una incorporación de etiqueta también puede compensarse mediante el grado redundancia de secuenciación. La probabilidad de una incorporación errónea, en este caso, se trata de manera estadística.

La formación de imágenes y localización de moléculas individuales, principalmente mediante fluorescencia, es familiar para los expertos en la técnica (véase Trabesinger, W., *et al.*, *Anal Chem.*, 1999. 71(1): págs. 279-83, Harms, G., *et al.*, *Biophys. J.*, 1999. 177: págs. 2864-2870, Deschryver, F., *Pure & Appl. Chem*, 1998. 70: págs. 2147-2156., Bartko, A. y R. Dickson, *J Phys Chem B*, 1999. 103: págs. 11237-11241). Por tanto, se generan fácilmente archivos de datos que contienen información referente a la ubicación y tipo de etiqueta. En una realización de esta invención, el análisis de los datos de secuencia se realiza al final del procedimiento de secuenciación y después de haber adquirido todos los datos de secuenciación. Estos datos, en uno o más archivos, pueden analizarse para determinar las ubicaciones de moldes e identificar cualquier indicador unido en estas posiciones. Tales datos se someten entonces a un análisis secundario para construir secuencias para todos los moldes localizados.

Preferiblemente, se realizan de manera simultánea ciclos de reacción de secuenciación y análisis de datos. En este caso, los datos generados de cada ciclo se analizan para localizar moléculas indicadoras, estas ubicaciones se correlacionan entonces con las ubicaciones de los moldes. Las secuencias para cada molde localizado pueden construirse en cada ciclo sucesivo.

El procedimiento preferido de realización de la invención se describirá ahora haciendo referencia a la figura 1.

En el sistema ilustrado en la figura 1, las moléculas que deben secuenciarse se han fijado sobre fases sólidas mediante procedimientos convencionales tal como se describe en la técnica. (Revisado en Schena (ed.), *DNA Microarrays: A practical approach*, Oxford University Press (1999) ISBN: 0199637768). El molde, unido a una superficie sólida tal como un portaobjetos de microscopio, se marca de manera que puede determinarse su posición cuando se observa el portaobjetos a través de un explorador de microscopio fluorescente, por ejemplo. En la etapa 10, se localiza en primer lugar un molde relevante.

ES 2 281 426 T3

Ahora se realizan, etapa 12, las reacciones de secuenciación que implican la incorporación de bases que puede realizarse extendiendo el molde con una enzima polimerasa o ligando un oligonucleótido marcado con una ligasa.

5 Tal como se describió anteriormente, la etapa de secuenciación dará como resultado la unión de una molécula indicadora a la primera base en la secuencia del molde, y la detección del resto indicador que se incorpora permite asignar la identidad de la base que debe asignarse, etapa 14. La siguiente etapa, etapa 16, es correlacionar las ubicaciones de base y molde; en este primer ciclo ésta es una etapa trivial. Entonces se realiza una determinación sobre si la molécula de molde está asociada con un indicador. Es decir, en la etapa 18 se realiza una prueba sobre si el molde objeto tiene un indicador o no. Si tras la operación de secuenciación el molde está asociado con un indicador, el procedimiento
10 avanza hacia la etapa 20. Aquí, se realiza una prueba para determinar si el indicador procede de un ciclo previo. Si no es así, entonces se identifica y se asigna una nueva base, etapa 22. Por tanto, se ha identificado correctamente la base y todo está bien.

15 Entonces el procedimiento avanza hacia la etapa 24 en la que se realiza una prueba sobre si hay más moldes. En ese caso, el procedimiento se repite desde la etapa 18.

20 Si en la etapa 20 se determina que el indicador asociado a la base procede de un ciclo anterior, entonces no se asigna ninguna base, etapa 26, y el procedimiento avanza directamente hacia la etapa 24 y hacia el siguiente molde, si lo hay.

25 Si en la etapa 18 se encuentra que el molde no tiene un indicador, en la etapa 50 se realiza una comprobación sobre si el señalizador de apareamiento erróneo está encendido. El señalizador de apareamiento erróneo se activa cuando el número de detenciones consecutivas supera el máximo predeterminado, según una prueba realizada en la etapa 30. Si el señalizador de apareamiento erróneo no está encendido, el procedimiento avanza hacia la etapa 28, y se inserta una detención D en la secuencia. Además, se aumenta en uno un contador de detenciones, que monitoriza el número de detenciones consecutivas que se producen. Se realiza una prueba en la etapa 30 para determinar si el número de detenciones consecutivas supera un umbral o valor máximo predeterminado. En caso contrario, el procedimiento avanza hacia la etapa 24 dejando la detención en la secuencia. Si el número de detenciones consecutivas supera el máximo predeterminado, entonces se puntúa la base anterior como con apareamiento erróneo y se activa el señalizador
30 de apareamiento erróneo, etapa 32, y entonces el procedimiento avanza hacia la etapa 24.

35 El indicador de detención sirve para la función de proporcionar una indicación de una base desconocida. Puede resultar ser una cualquiera de las bases A, C, G y T, o de hecho puede resultar no ser ninguna base. Al prever la posibilidad de una base desconocida, no se descarta completamente la información para ese molde. En vez de eso, todavía puede usarse, por ejemplo con respecto a una secuencia de referencia, tal como se describe en los ejemplos siguientes.

40 Si en la etapa 20 se determina que el indicador procede de un ciclo anterior, en la etapa 52 se realiza una comprobación sobre si el señalizador de apareamiento erróneo está encendido. Si el señalizador de apareamiento erróneo no está encendido, entonces el procedimiento avanza hacia la etapa 22 y se asigna una base. Entonces el procedimiento avanza hacia la etapa 24 para determinar si hay otro molde para tratar.

45 Si el señalizador de apareamiento erróneo está encendido, en la etapa 54 se sustituye la base previamente asignada con un código IUB que representa todas las otras bases excepto la que estaba apareada erróneamente. Esto es porque si la base anterior estaba marcada como "C" pero se sabe que esto era un apareamiento erróneo, queda claro que la base es A, G o T.

50 Cuando no hay más moldes, la prueba en la etapa 24 tiene el resultado NO, y el procedimiento avanza hacia la etapa 34, en la que se realiza una determinación sobre si hay más ciclos que deben completarse, es decir, si hay más bases para esa molécula. Si las hay, el procedimiento avanza hacia los datos para el siguiente ciclo, etapa 36, tras lo cual el procedimiento avanza de nuevo desde la etapa 16, con correlación de las ubicaciones de molde y base.

55 Eventualmente, la prueba en la etapa 34 tendrá el resultado NO, y esto conduce al final del procedimiento, etapa 38.

Pueden aplicarse tratamientos posteriores a la secuencia tal como se produce por el sistema de la figura 1, por ejemplo para comparar la secuencia encontrada mediante el procedimiento con una secuencia de referencia. A continuación se describen ejemplos de esto.

60 Las etapas mostradas en la figura 1, posteriores a las etapas 10 a 14 que implican reacciones químicas, se realizan en un ordenador digital tal como un ordenador personal (PC). Se muestran dos ejemplos en más detalle a modo de pseudocódigo en el apéndice de esta memoria descriptiva. El primer pseudocódigo supone que se requieren los nucleótidos por una mezcla de las cuatro bases A, C, G, y T, y el segundo pseudocódigo es para su uso cuando las cuatro bases se usan por separado en la secuencia.

65 La presente invención tiene muchas aplicaciones, algunas de las cuales se facilitan en el presente documento. Por ejemplo, puede determinarse la secuencia de genomas de ADN y ARN usando este procedimiento. Además, pueden identificarse las variaciones de secuencia en regiones o genomas enteros, representaciones de ARNm de regiones o

ES 2 281 426 T3

genomas enteros o en representaciones generadas artificialmente de un genoma (por ejemplo, productos de PCR de regiones de un genoma) que resultan de sustituciones, delecciones o inserciones de una o más bases.

La presente invención tiene aplicación en la determinación de haplotipo (determinación de diferencias de secuencia entre pares de cromosomas en un individuo) y también en el análisis de la expresión de ARNm cuantitativa, por ejemplo para comparar niveles de expresión de ARNm entre muestras derivadas de distintos tipos celulares (tejidos) o células tratadas de manera diferente. Esta técnica también puede aplicarse para identificar secuencias derivadas de genomas de patógenos para su uso en la detección e identificación de patógenos.

Ahora se facilitan ejemplos del modo en que se tratan secuencias específicas por el sistema de tal manera que se reducen errores en la secuencia determinada.

Ejemplo 1

La siguiente secuencia se obtiene a partir de una reacción de secuenciación:

GATCGGCTGACCATGGAC1

en la que 1 indica que se ha incorporado una T (y 2=C, 3=A, 4=G).

Un fallo de extensión adicional para el número de ciclos umbral da como resultado el marcaje de la secuencia para indicar que se ha incorporado erróneamente una base antes del número de detenciones umbral. Aquí, un 1 (uno) indica que se ha incorporado una T antes de un número de detenciones por encima del nivel de umbral determinado y por tanto es probable que se haya incorporado erróneamente. Por tanto, puede descartarse la secuencia. Haciendo referencia a la figura 1, el procedimiento sigue la ruta 28, 30 durante un número predeterminado de etapas, hasta que el resultado en la etapa 30 es SÍ, y la base anterior se marca como con apareamiento erróneo en la etapa 32. En lugar de un 1, para las otras bases se usa 2, 3 ó 4, indicando 2 una C, indicando 3 una A, e indicando 4 una G.

Ejemplo 2

Las siguientes secuencias se obtienen de una reacción de secuenciación. La primera es una secuencia recién determinada y la segunda es una secuencia de referencia:

GATCGGCTGACCATGGACC1CTGACAGT

GATCGGCTGACCATGGACCTCTGACAGT

Detener durante más tiempo que el número de ciclos umbral marca un 1 para T como una incorporación errónea. En este caso, la secuenciación se ha reanudado tras el número de secuencias umbral. Cuando la secuencia obtenida se compara con la secuencia de referencia, la alineación de secuencias demuestra una alineación T.1 en la posición detenida. Por tanto, puede descartarse como una diferencia de base real con la secuencia de referencia. La alineación de secuencias representa una fase adicional al tratamiento ilustrado en la figura 1.

Ejemplo 3

Cuando se encuentra una detención durante la secuenciación, su posición se marca como D. Si se obtienen las siguientes secuencias nueva y de referencia:

GATCGGCTGACCATGGADCCTCTGACAGT

GATCGGCTGACCATGGACCTCTGACAGT

la alineación de secuencias con la secuencia de referencia en presencia o ausencia de un hueco en la posición marcada con una D revela que era una detención. Por tanto, toda la secuencia es contigua y útil. La alineación de secuencias representa de nuevo una fase adicional al tratamiento ilustrado en la figura 1.

Ejemplo 4

Se obtiene la siguiente secuencia en una reacción de secuenciación:

GATCGGCTGACCATGGDCCTCTGACAGT

GATCGGCTGACCATGGACCTCTGACAGT

La posición marcada como D es la incorporación de una base con un indicador que ha fallado. La alineación de secuencias con una secuencia de referencia en presencia o ausencia de un hueco en la posición marcada revela que esto representa un hueco en la secuencia. La secuencia extraída sigue siendo útil. En este caso, la D puede sustituirse con una "N" para representar un hueco en la secuencia. La alineación de secuencias representa de nuevo una fase adicional al tratamiento ilustrado en la figura 1.

APÉNDICE

Primer pseudocódigo

Ejemplo de un pseudocódigo para el montaje de secuencia tras completar las reacciones de secuenciación.

Main ()

{

Localizar moldes ();

For (; número de ciclos que deben analizarse;);

{

Correlacionar indicadores con ubicaciones de moldes ()

While (haya moldes)

{

si (la ubicación del molde no tiene un indicador)

{

 aumentar el contador de detenciones;

 if (contador de detenciones > umbral) marcar la base anterior como apareamiento erróneo;

}

 else if (el indicador procede del ciclo anterior) descartar;

 else identificar y asignar base a molde;

 avanzar hacia el siguiente molde;

}

if (deben analizarse más ciclos) avanzar hacia datos para el siguiente ciclo de secuenciación;

else return;

}

}

Segundo pseudocódigo

Pseudocódigo para la secuenciación de bases únicas secuenciales

5

Main ()

{

10

Ubicar moldes ();

While (haya ciclos)

{

15

leer los datos para el ciclo

For (;cuatro bases;)

{

20

Correlacionar indicadores con ubicaciones de molde ()

While (haya moldes)

{

25

if (el indicador procede de un ciclo anterior) descartar;

else identificar y asignar base a molde;

marcar molde como extendido ();

siguiente molde

30

}

aumentar el marcador de detenciones para todos los moldes no
marcados como extendido ();

35

while (moldes detenidos)

{

40

if (número de detenciones ha alcanzado el umbral) marcar la base
anterior como incorporada erróneamente

}

avanzar hacia el siguiente ciclo ();

45

}

Secuencia de salida para su análisis ();

}

50

55

60

65

REIVINDICACIONES

1. Un procedimiento para determinar la secuencia de un ácido nucleico en el que se añaden nucleótidos secuencialmente a un cebador en un molde, estando dicho molde unido a un soporte sólido, que comprende las etapas de
 - a) hibridar un cebador al molde, comprendiendo dicho molde el ácido nucleico para el que se requiere la secuencia, para formar un complejo de molde-cebador,
 - b) asignar uno o más nucleótidos marcados, que comprenden uno o más de A, C, G o T, y extender el cebador mediante una ADN polimerasa o una ligasa intentando añadir el nucleótido marcado único, proporcionándose el marcaje mediante un resto indicador,
 - c) determinar el tipo o identidad del nucleótido marcado con indicador añadido al cebador, de manera que se identifica un nucleótido en la secuencia del molde,
 - d) eliminar o neutralizar el resto indicador,
 - e) repetir las etapas b) a d) secuencialmente en cada ciclo siguiente y registrar el orden de incorporación de nucleótidos marcados con indicador,
 - f) determinar si un resto indicador puede detectarse en la etapa c) de cada ciclo,
 - g) si el resto indicador no puede detectarse en la etapa c) de cada ciclo, introduciendo un indicador de detención en la secuencia para proporcionar una indicación de una base desconocida,
 - h) si el resto indicador no puede detectarse, determinar si el número de indicadores de detención consecutivos supera un umbral predeterminado y, en ese caso, asignar un señalizador de apareamiento erróneo,
 - i) si el resto indicador puede detectarse en la etapa c) y no se ha identificado en la iteración anterior de las etapas b) a d) o si no se detecta señalizador de apareamiento erróneo, asignar la base correspondiente a la secuencia,
 - j) si el resto indicador puede detectarse en la etapa c) y se ha identificado en la iteración anterior de las etapas b) a d) y si se detecta el señalizador de apareamiento erróneo, sustituir la base previamente asignada con un código IUB para todas las demás bases, y
 - k) si no puede detectarse el resto indicador en la etapa c) y si se observa ausencia del resto indicador para menos de un número umbral predeterminado de una o más iteraciones de las etapas b) a d) entonces el nucleótido correspondiente al indicador en la iteración actual de las etapas b) a d) se incorpora a la secuencia de manera contigua a la base previamente incorporada.
2. Procedimiento según la reivindicación 1, en el que se anota un marcador posicional en la secuencia entre el nucleótido incorporado en la iteración actual de b) a d) y el nucleótido que lo precede para indicar un posible hueco en la secuencia.
3. Procedimiento según la reivindicación 1 ó 2, que comprende adicionalmente la etapa de determinar la posición del molde detectando la ubicación del resto indicador.
4. Procedimiento según la reivindicación 3, en el que la ubicación del molde se determina detectando el resto indicador unido al molde.
5. Procedimiento según la reivindicación 3, en el que la ubicación del molde se determina detectando un resto indicador unido al cebador.
6. Procedimiento según la reivindicación 1, en el que cualquier adición posterior se considera una continuación de la secuencia.
7. Procedimiento según la reivindicación 6 en el que se anota un marcador posicional en la secuencia entre el nucleótido incorporado en la iteración actual de b) a d) y el nucleótido que lo precede para indicar un posible hueco en la secuencia.
8. Un medio de código de programa informático para realizar todas las etapas según una cualquiera de las reivindicaciones 1 a 7, cuando se ejecuta dicho programa en un ordenador.

