(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 964 351**

(51) Int. Cl.:

G16B 50/50 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **11.09.2020 PCT/US2020/050584**

(87) Fecha y número de publicación internacional: **18.03.2021 WO21051019**

(96) Fecha de presentación y número de la solicitud europea: **11.09.2020 E 20788695 (3)**

(97) Fecha y número de publicación de la concesión europea: **06.09.2023 EP 4029023**

(54) Título: **Método para la compresión de datos de secuencia genómica**

(30) Prioridad:

11.09.2019 US 201916567211

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
05.04.2024

(73) Titular/es:

ILLUMINA, INC. (100.0%)
5200 Illumina Way
San Diego, CA 92122, US

(72) Inventor/es:

RIZK, GUILLAUME ALEXANDRE PASCAL

(74) Agente/Representante:

DEL VALLE VALIENTE, Sonia

DESCRIPCIÓN

Método para la compresión de datos de secuencia genómica

Campo técnico

El campo se refiere generalmente a los métodos de representación de los datos de secuenciación genómica producidos por una máquina de secuenciación, y más particularmente a los métodos implementados por ordenador para la compresión de tales datos de secuenciación genómica. Esta descripción proporciona un método de compresión basado en referencia que permite una compresión y descompresión rápidas mientras que no provoca ninguna pérdida de información, y que tiene una alta razón de compresión.

Antecedentes

Las máquinas de secuenciación de próxima generación producen ahora grandes cantidades de datos de secuenciación a un precio asequible. Los sistemas recientes producen en una sola ejecución de 36 h más de 6 mil millones de secuencias largas de 150 nucleótidos, lo suficiente para la secuenciación de 20 genomas humanos completos. Esto abre muchas nuevas perspectivas para el diagnóstico de enfermedades genéticas y para el desarrollo de la medicina personalizada, con el objetivo de adaptar el tratamiento basándose en las especificidades genómicas de las personas.

Sin embargo, esto también viene con nuevos desafíos, en particular el coste relacionado con el almacenamiento de grandes cantidades de datos. El formato de archivo más usado para los datos de secuencia sin procesar (no alineados) es el formato FASTQ, que mantiene datos de secuencia (cadena de nucleótidos A, C, T, G también denominada lectura), valores de calidad (probabilidades de que la plataforma de secuenciación elaboró un error de secuenciación para cada nucleótido) y nombres de secuencia. Este es un archivo de texto ASCII sin formato, habitualmente comprimido con el esquema de compresión de texto de uso general LZ (esquema de Lemel-Ziv, implementado en el software gzip). Sin embargo, el uso de tales métodos de compresión viene con varios problemas:

- razón de compresión baja porque la redundancia de los datos no se utiliza completamente
- compresión y descompresión lenta

También existen métodos de compresión especializados en la codificación FASTQ, divididos en métodos de referencia o no basados en referencia. Sin embargo, ninguno de ellos es completamente satisfactorio, ya que a) los métodos basados en referencia tienen buenas razones de compresión pero son lentos, b) los métodos no basados en referencia son más rápidos pero tienen razones de compresión más bajas. Un ejemplo de un método no basado en referencia de tipo se proporciona por el software SPRING, que es un compresor sin referencia para archivos FASTQ (dirección de Internet: github.com/shubhamchandak94/SPRING). Sin embargo, el método de compresión proporcionado por el software SPRING tiene una razón de compresión baja.

Entre los métodos de compresión basados en referencia, se han propuesto algunos métodos que usan alineaciones de secuencia y tienen como objetivo ser más rápidos con buenas razones de compresión. Sin embargo, tales métodos adolecen de varios problemas, notablemente un problema importante es que no son completamente sin pérdidas. Un método de compresión basado en referencia conocido de este tipo se describe, por ejemplo, en el documento de patente WO 2018/068829 A1. En el método descrito, después de haberse alineado con una o más secuencias de referencia, las secuencias de nucleótidos se clasifican según los grados de precisión coincidentes (creando de ese modo clases de lecturas alineadas) y luego se codifican como una multiplicidad de capas de elementos sintácticos, usando diferentes modelos de origen y codificadores de entropía para cada capa en la que se dividen los datos. Las clases de datos se codifican así por separado y se estructuran en diferentes capas de elementos sintácticos, comprendiendo cada capa descriptores que representan unívocamente las lecturas clasificadas y alineadas de dicha capa. El método está destinado a obtener distintas fuentes de información con entropía de información reducida, permitiendo de ese modo un aumento en el rendimiento de compresión, así como un acceso selectivo a clases específicas de datos comprimidos. Sin embargo, un método de compresión de este tipo reordena las lecturas en un orden que es diferente del obtenido al final del paso de alineación de lectura (es decir, las lecturas se reordenan según sus clases). Luego, parte de la información se pierde en el proceso de compresión, en particular el orden de la secuencia inicial. Por tanto, la reproducibilidad de algunos resultados de análisis puede verse afectada, porque algún software de análisis posterior puede depender del orden de las lecturas. Además, descomprimir los datos en un orden que es diferente del orden inicial de las lecturas hace que sea mucho más difícil comprobar que el archivo no comprimido es idéntico al archivo inicial. Además, un método de compresión de este tipo es relativamente lento, especialmente cuando se compara con los métodos de compresión no basados en referencia del estado de la técnica. La publicación de patente estadounidense número US 2015/227686 describe un método para alinear un genoma de referencia con una pluralidad de secuencias de ADN.

Resumen

Las características de las reivindicaciones independientes a continuación resuelven el problema de las soluciones existentes de la técnica anterior proporcionando un método para la compresión de los datos de la secuencia del genoma. En un aspecto, un método implementado por ordenador para la compresión de datos de secuencia genómica producidos por una máquina de secuenciación, comprendiendo dichos datos de secuencia genómica lecturas de secuencias de nucleótidos o bases que se han alineado con respecto a una secuencia de referencia, creando de ese modo lecturas alineadas, almacenándose dichas lecturas alineadas como una lista de lecturas en un archivo inicial, que comprende los pasos de

- para cada lectura alineada, determinar si dicha lectura se correlaciona perfecta o imperfectamente con dicha secuencia de referencia o si dicha lectura no se correlaciona con dicha secuencia de referencia,

- codificar las lecturas según dicha determinación, en donde las lecturas que se determinan que están perfectamente correlacionadas se codifican según un primer proceso de codificación y las lecturas que se determinan que no están correlacionadas se codifican según un segundo proceso de codificación,

- en donde el paso de determinación comprende comparar, para cada lectura imperfectamente correlacionada, el número de apareamientos erróneos entre dicha lectura y dicha secuencia de referencia con un valor umbral,

- en donde, en el paso de codificación, las lecturas que se determinan que están imperfectamente correlacionadas se codifican según el segundo proceso de codificación o un tercer proceso de codificación, las lecturas imperfectamente correlacionadas se codifican según el segundo proceso de codificación cuando dicho número de apareamientos erróneos es mayor que el valor umbral, las lecturas imperfectamente correlacionadas se codifican según el tercer proceso de codificación cuando dicho número de apareamientos erróneos es menor que el valor umbral,

- en donde, en dicho segundo proceso de codificación, cada nucleótido o base de la lectura se codifica individualmente,

- en donde dichos procesos de codificación primero y tercero comprenden conjuntos distintos de descriptores, representando cada conjunto de descriptores unívocamente las lecturas asociadas al proceso de codificación correspondiente, siendo cada uno de dichos procesos de codificación primero y tercero un proceso de codificación por entropía de fuente de información reducida.

La invención supera las desventajas de los métodos de compresión anteriores permitiendo una rápida compresión y descompresión sin provocar pérdida de información, y proporcionando una alta razón de compresión. Más particularmente, la invención se centra en codificar los casos más frecuentes de la manera más compacta, incluso si esto significa adoptar modos de codificación degradados para los casos raros menos frecuentes. Esto conduce a un gran aumento en el rendimiento de compresión. Además, debido al formato de representación de información genómica que se usa en la invención, la compresión realizada mediante el método de la invención es más rápida. Por último, pero no menos importante, el método según la invención mantiene el orden inicial de las lecturas como tal y no reordena las lecturas según sus clases. En consecuencia, no se pierde información durante el proceso, lo que permite un análisis posterior más fácil, así como comprobaciones de conformidad eficientes después del paso de descompresión.

Estas y otras características y ventajas de la presente invención resultarán más evidentes a partir de los dibujos adjuntos y de la siguiente descripción detallada. Además, aunque puede hacerse referencia a los umbrales en el presente documento como superados o no superados, se entiende que tales umbrales pueden emplearse conceptualmente de modo que se determine si tal umbral se satisface, se cumple o se detecta de otro modo, independientemente de si los números o valores usados para implementar esas evaluaciones de umbral se describen usando valores positivos o negativos.

Según un aspecto innovador de la presente memoria, se describe un método para comprimir datos de secuencia genómica. En un aspecto, el método puede incluir el rendimiento de una o más operaciones a través de la ejecución de instrucciones de software mediante uno o más ordenadores, donde las operaciones incluyen obtener, mediante el uno o más ordenadores, un registro de lectura, determinar, mediante el uno o más ordenadores, si el registro de lectura corresponde a una lectura que se correlaciona perfectamente con una secuencia de referencia o se correlaciona imperfectamente con la secuencia de referencia, basándose en la determinación, mediante el de uno o más ordenadores, de que el registro de lectura corresponde a una lectura que se correlaciona imperfectamente con la secuencia de referencia, determinar, mediante el uno o más ordenadores, si un número de apareamientos erróneos de la lectura imperfectamente correlacionada satisface un número umbral predeterminado de apareamientos erróneos, y basándose en la determinación de que el número de apareamientos erróneos satisface el número umbral predeterminado de apareamientos erróneos, codificar, mediante el uno o más ordenadores, cada apareamiento erróneo de la lectura imperfectamente correlacionada en un registro que tiene un tamaño de 1 byte.

Otros aspectos incluyen sistemas, aparatos y programas informáticos correspondientes para realizar las acciones de los métodos tal como se describe en el presente documento como se define mediante instrucciones codificadas en dispositivos de almacenamiento legibles por ordenador.

Estas y otras versiones pueden incluir opcionalmente una o más de las siguientes características. Por ejemplo, en algunas implementaciones, determinar, mediante uno o más ordenadores, si un número de apareamientos erróneos de la lectura imperfectamente correlacionada satisface un número umbral predeterminado de apareamientos erróneos puede incluir determinar, mediante el uno o más ordenadores, si el número de apareamientos erróneos de la lectura imperfectamente correlacionada es mayor que el número umbral predeterminado de apareamientos erróneos.

En algunas implementaciones, cada registro de lectura puede incluir datos que indican una posición de inicio absoluta de una lectura alineada con respecto a la secuencia de referencia, datos que indican una longitud de la lectura, datos que indican si la lectura está perfectamente correlacionada o imperfectamente correlacionada, datos que indican un número de apareamientos erróneos identificados en la lectura, y datos que indican una posición relativa de cada uno de dichos posibles apareamientos erróneos en la lectura.

En algunas implementaciones, codificar cada apareamiento erróneo de la lectura imperfectamente correlacionada en un registro que tiene un tamaño de 1 byte comprende para cada apareamiento erróneo particular: codificar, mediante el uno o más ordenadores, unos primeros dos bits del byte para incluir datos que representan un nucleótido o base alternativa presente en la lectura en lugar de un nucleótido o base de referencia correspondiente en la secuencia de referencia, y codificar, mediante el uno o más ordenadores, los seis bits restantes del byte para incluir datos que representan una posición del apareamiento erróneo en la secuencia de referencia, calculándose dicha posición como un desplazamiento de un apareamiento erróneo previo de la lectura.

En algunas implementaciones, el método puede incluir además determinar, mediante uno o más ordenadores, si el desplazamiento es mayor que un valor codificable máximo, y basándose en la determinación de que el desplazamiento es mayor que el valor codificado máximo, insertar, mediante uno o más ordenadores, al menos un falso apareamiento erróneo entre el apareamiento erróneo particular y el apareamiento erróneo previo.

En algunas implementaciones el método puede incluir además basándose en la determinación de que el número de apareamientos erróneos no satisface el número umbral predeterminado de apareamientos erróneos, codificar, mediante uno o más ordenadores, una lista de posiciones de la secuencia de referencia correspondiente a una posición de cada uno de los apareamientos erróneos con respecto a la secuencia de referencia usando un proceso de codificación por entropía de información reducida.

En algunas implementaciones, el método puede incluir además adicionalmente, basándose en la determinación de que el registro de lectura corresponde a una lectura que se correlaciona perfectamente con la secuencia de referencia, codificar, mediante uno o más ordenadores, al menos una parte del registro de lectura usando codificación por entropía de información reducida.

En algunas implementaciones, el uno o más ordenadores pueden incluir uno o más procesadores de hardware.

En algunas implementaciones, el uno o más procesadores de hardware pueden incluir una o más matrices de puertas programables en campo (FPGA).

En algunas implementaciones, el método para comprimir datos de secuencia genómica puede realizarse mediante uno o más procesadores de hardware. En tales implementaciones, los procesadores de hardware pueden incluir un conjunto de circuitos de procesamiento de hardware que se configura para realizar una o más operaciones. En un aspecto, las operaciones pueden incluir obtener, mediante el conjunto de circuitos de procesamiento de hardware, un registro de lectura, determinar, mediante el conjunto de circuitos de procesamiento de hardware, si el registro de lectura corresponde a una lectura que se correlaciona perfectamente con una secuencia de referencia o se correlaciona imperfectamente con la secuencia de referencia, basándose en la determinación, mediante el conjunto de circuitos de procesamiento de hardware, de que el registro de lectura corresponde a una lectura que se correlaciona imperfectamente con la secuencia de referencia, determinar, mediante el uno o más ordenadores, si un número de apareamientos erróneos de la lectura imperfectamente correlacionada satisface un número umbral predeterminado de apareamientos erróneos, y basándose en la determinación de que el número de apareamientos erróneos satisface el número umbral predeterminado de apareamientos erróneos, codificar, mediante el conjunto de circuitos de procesamiento de hardware, cada apareamiento erróneo de la lectura imperfectamente correlacionada en un registro que tiene un tamaño de 1 byte.

Estas y otras características y ventajas de la presente invención resultarán más evidentes a partir de los dibujos adjuntos y de la siguiente descripción detallada.

En algunas implementaciones, cada registro de lectura puede incluir: datos que indican una posición de inicio absoluta de la lectura alineada con respecto a la secuencia de referencia, datos que indican una longitud de la lectura, datos

que indican si la lectura está perfectamente correlacionada o imperfectamente correlacionada, datos que indican un número de apareamientos erróneos identificados en la lectura, y datos que indican una posición relativa de dichos posibles apareamientos erróneos en la lectura.

En algunas implementaciones, determinar, mediante el conjunto de circuitos de procesamiento de hardware, si un número de apareamientos erróneos de la lectura imperfectamente correlacionada satisface un número umbral predeterminado de apareamientos erróneos puede incluir determinar, mediante el conjunto de circuitos de procesamiento de hardware, si el número de apareamientos erróneos de la lectura imperfectamente correlacionada es mayor que el número umbral predeterminado de apareamientos erróneos.

En algunas implementaciones, codificar cada apareamiento erróneo de la lectura imperfectamente correlacionada en un registro que tiene un tamaño de 1 byte puede incluir para cada apareamiento erróneo particular, codificar, mediante el conjunto de circuitos de procesamiento de hardware, unos primeros dos bits del byte para incluir datos que representan un nucleótido o base alternativa presente en la lectura en lugar de un nucleótido o base de referencia correspondiente en la secuencia de referencia, y codificar, mediante el conjunto de circuitos de procesamiento hardware, los seis bits restantes del byte para incluir datos que representan una posición del apareamiento erróneo en la secuencia de referencia, calculándose dicha posición como un desplazamiento de un apareamiento erróneo previo de la lectura.

En algunas implementaciones, el conjunto de circuitos de procesador de hardware se configura además para realizar operaciones que incluyen determinar, mediante el conjunto de circuitos de procesamiento de hardware, si el desplazamiento es mayor que un valor codificable máximo, y basándose en la determinación de que el desplazamiento es mayor que el valor codificado máximo, insertar, mediante el conjunto de circuitos de hardware, al menos un falso apareamiento erróneo entre el apareamiento erróneo particular y el apareamiento erróneo previo.

En algunas implementaciones el conjunto de circuitos de procesador de hardware se configura además para realizar operaciones que incluyen basándose en la determinación de que el número de apareamientos erróneos no satisface el número umbral predeterminado de apareamientos erróneos, codificar, mediante el conjunto de circuitos de procesamiento de hardware, una lista de posiciones de la secuencia de referencia correspondiente a una posición de cada uno de los apareamientos erróneos con respecto a la secuencia de referencia usando un proceso de codificación por entropía de información reducida.

En algunas implementaciones, el conjunto de circuitos de procesador de hardware se configura además para realizar operaciones que incluyen, basándose en la determinación de que el registro de lectura corresponde a una lectura que se correlaciona perfectamente con la secuencia de referencia, codificar, mediante el conjunto de circuitos de procesamiento de hardware, al menos una parte del registro de lectura usando codificación por entropía de información reducida.

En algunas implementaciones, el conjunto de circuitos de procesamiento de hardware comprende una o más matrices de puertas programables en campo (FPGA).

Según otro aspecto innovador de la presente memoria, un método implementado por ordenador para la compresión de datos de secuencia genómica producidos por una máquina de secuenciación, comprendiendo dichos datos de secuencia genómica lecturas de secuencias de nucleótidos o bases que se han alineado con respecto a una secuencia de referencia, creando de ese modo lecturas alineadas, almacenándose dichas lecturas alineadas como una lista de lecturas en un archivo inicial. En un aspecto, el método puede incluir acciones de cada lectura alineada, determinar si dicha lectura se correlaciona perfecta o imperfectamente con dicha secuencia de referencia o si dicha lectura no se correlaciona con dicha secuencia de referencia, codificar las lecturas según dicha determinación, en donde las lecturas que se determina que están perfectamente correlacionadas se codifican según un primer proceso de codificación y las lecturas que se determinan que no están correlacionadas se codifican según un segundo proceso de codificación, en donde el paso de determinación comprende comparar, para cada lectura imperfectamente correlacionada, el número de apareamientos erróneos entre dicha lectura y dicha secuencia de referencia con un valor umbral, en donde, en el paso de codificación, las lecturas que se determinan que están imperfectamente correlacionadas se codifican según el segundo proceso de codificación o un tercer proceso de codificación, codificándose las lecturas imperfectamente correlacionadas según el segundo proceso de codificación cuando dicho número de apareamientos erróneos es mayor que el valor umbral, codificándose las lecturas imperfectamente correlacionadas según el tercer proceso de codificación cuando dicho número de apareamientos erróneos es menor que el valor umbral, en donde, en dicho segundo proceso de codificación, se codifica individualmente cada nucleótido o base de la lectura, en donde dichos primer y tercer procesos de codificación comprenden distintos conjuntos de descriptores, representando cada conjunto de descriptores unívocamente las lecturas asociadas a un proceso de codificación correspondiente, siendo cada uno de dichos primer y tercer procesos de codificación un proceso de codificación por entropía de fuente de información reducida.

Otros aspectos incluyen sistemas, aparatos y programas informáticos correspondientes para realizar las acciones de los métodos tal como se describe en el presente documento como se define mediante instrucciones codificadas en dispositivos de almacenamiento legibles por ordenador.

Estas y otras versiones pueden incluir opcionalmente una o más de las siguientes características. Por ejemplo, en algunas implementaciones, el paso de determinación puede incluir, cuando se determina que una lectura que se correlaciona imperfectamente con la secuencia de referencia y tiene un número de apareamientos erróneos inferiores al valor umbral, una determinación adicional en cuanto a si la lectura se correlaciona global o localmente con dicha secuencia de referencia, y en donde el tercer proceso de codificación comprende un primer subproceso de codificación y un segundo subproceso de codificación, las lecturas que se determina que están globalmente correlacionadas se codifican según el primer subproceso de codificación, las lecturas que se determinan que están localmente correlacionadas se codifican según el segundo subproceso de codificación, dichos primer y segundo subprocesos de codificación que comprenden conjuntos distintos de descriptores, representando cada conjunto de descriptores unívocamente las lecturas asociadas a un subproceso de codificación correspondiente.

En algunas implementaciones, dichos descriptores de dicho primer subproceso de codificación pueden incluir una posición de inicio de alineación en la secuencia de referencia, una longitud de lectura y una lista de apareamientos erróneos por sustituciones de símbolos, y en donde dichos descriptores de dicho segundo subproceso de codificación comprenden una posición de inicio de alineación local en la secuencia de referencia, una longitud de lectura, una lista de apareamientos erróneos por sustituciones de símbolos, y una longitud de las partes recortadas de la lectura que no son parte de la alineación.

En algunas implementaciones, en el paso de codificación, las partes recortadas de una lectura que se va a codificar según el segundo subproceso de codificación se concatenan, codificándose cada nucleótido o base de dichas partes recortadas individualmente.

En algunas implementaciones, en el paso de codificación, cada apareamiento erróneo de una lectura imperfectamente correlacionada se codifica en 1 byte.

En algunas implementaciones, en el paso de codificación, cada apareamiento erróneo de una lectura imperfectamente correlacionada se codifica de la siguiente manera: dos primeros bits del byte se usan para codificar un nucleótido o base alternativa presente en la lectura en lugar de un nucleótido o base de referencia correspondiente en la secuencia de referencia, y los seis últimos bits del byte se usan para codificar una posición del apareamiento erróneo en la secuencia de referencia, calculándose dicha posición como un desplazamiento de un apareamiento erróneo previo de la lectura.

En algunas implementaciones, en el paso de codificación, si el desplazamiento calculado entre un apareamiento erróneo determinado y el apareamiento erróneo previo es mayor que un valor codificable máximo, entonces se inserta al menos un falso apareamiento erróneo entre dichos dos apareamientos erróneos hasta que cada desplazamiento entre cada una de dichos apareamientos erróneos y dicha al menos un falso apareamiento erróneo es menor que dicho valor codificable máximo, definiéndose un falso apareamiento erróneo como un apareamiento erróneo para el cual bits del byte usado para codificar el apareamiento erróneo o para codificar un nucleótido o base que es igual al correspondiente nucleótido o base de referencia en la secuencia de referencia.

En algunas implementaciones, un paso inicial de dividir la lista de lecturas en bloques de lecturas, comenzando cada bloque con un encabezado que contiene información necesaria para decodificar el bloque, donde dicho método de compresión se realiza bloque por bloque.

En algunas implementaciones, los bloques de lecturas tienen el mismo tamaño de bloque.

En algunas implementaciones, un paso final de proporcionar un archivo comprimido que comprende una lista de lecturas codificadas, almacenándose dichas lecturas codificadas en el archivo comprimido en el mismo orden que el de las lecturas almacenadas en el archivo inicial.

En algunas implementaciones, dicho valor umbral es igual a 31.

En algunas implementaciones, para cada lectura alineada, un paso de determinar si dicha lectura comprende al menos un apareamiento erróneo correspondiente a un caso en el que la máquina de secuenciación no pudo llamar a ninguna base o nucleótido.

En algunas implementaciones, para cada lectura que comprende al menos un apareamiento erróneo correspondiente a un caso en el que la máquina de secuenciación no pudo llamar a ninguna base o nucleótido, un paso de determinar el número de tales apareamientos erróneos y un paso de comparar dicho número con un valor umbral de referencia.

En algunas implementaciones, en el paso de codificación, si el número de tales apareamientos erróneos es mayor que el valor umbral de referencia, cada nucleótido o base de una lectura que se va a codificar según el segundo proceso de codificación se codifica individualmente en 4 bits, y, si el número de tales apareamientos erróneos es menor que el valor umbral de referencia, cada nucleótido o base de una lectura que va a codificarse según el segundo proceso de codificación se codifica individualmente en 2 bits y el paso de codificación comprende además codificar una lista de

posiciones a lo largo de la secuencia de referencia, correspondiendo dichas posiciones a las posiciones de tales apareamientos erróneos en la secuencia de referencia.

Breve descripción de los dibujos

La Figura 1 es un diagrama de flujo que muestra los pasos del método de compresión según la invención.

La Figura 2 es un diagrama que muestra un aparato para implementar los pasos del método de compresión según la invención.

La Figura 3 muestra un primer ejemplo de una lectura que se correlaciona globalmente con una secuencia de referencia.

La Figura 4 muestra un segundo ejemplo de una lectura que se correlaciona globalmente con una secuencia de referencia, en un caso en el que se ha de insertar un falso apareamiento erróneo.

Descripción detallada

Las secuencias genómicas a las que se hace referencia en esta invención incluyen, por ejemplo, y no como limitación, secuencias de nucleótidos, secuencias de ácido desoxirribonucleico (ADN), ácido ribonucleico (ARN) y secuencias de aminoácidos. Aunque la descripción en el presente documento tiene un detalle considerable con respecto a la información genómica en forma de una secuencia de nucleótidos, se entenderá que el método de compresión según la invención también puede implementarse para otras secuencias genómicas también, aunque con algunas variaciones, como entenderá un experto en la técnica.

La información de secuenciación genómica se genera mediante máquinas de secuenciación en forma de secuencias de nucleótidos (o, más generalmente, bases) representadas por cadenas de letras de un vocabulario definido. El vocabulario más pequeño está representado por cinco símbolos: {A, C, G, T, N} que representa los 4 tipos de nucleótidos presentes en el ADN, concretamente adenina, citosina, guanina y timina. En el ARN la timina se reemplaza por uracilo (U). N indica que la máquina de secuenciación no pudo llamar a ninguna base y, por tanto, la naturaleza real de la posición no está determinada.

Las secuencias de nucleótidos producidas por máquinas de secuenciación se denominan "lecturas". Las lecturas de secuencia pueden tener entre unas pocas docenas a varios miles de nucleótidos de longitud. Algunas tecnologías producen lecturas de secuencia en pares donde una lectura es de una cadena de ADN y la segunda es de la otra cadena. A lo largo de esta descripción, una "secuencia de referencia" es cualquier secuencia en la que las secuencias de nucleótidos o bases producidas por máquinas de secuenciación se alinean/correlacionan. Un ejemplo de tal secuencia de referencia podría ser realmente un genoma de referencia, es decir, una secuencia ensamblada por científicos como un ejemplo representativo de un conjunto de genes de especie. Sin embargo, una secuencia de referencia también podría consistir en una secuencia sintética concebida para mejorar simplemente la compresibilidad de las lecturas en vista de su posterior procesamiento. Las máquinas de secuenciación pueden introducir errores en las lecturas de secuencia, y notablemente un uso de un símbolo incorrecto (es decir, que representa un ácido nucleico diferente) para representar el ácido nucleico o la base realmente presente en la muestra secuenciada; esto se denomina habitualmente un error de sustitución o un "apareamiento erróneo".

La invención es un método de compresión basado en referencia que recibe lecturas de secuencias de nucleótidos o bases como entradas, tales lecturas se han alineado previamente con una secuencia de referencia, creando de ese modo lecturas alineadas. Luego se almacenan las lecturas alineadas como una lista de lecturas en un archivo inicial. La manera de alinear las lecturas y almacenarlas una vez alineadas en un archivo inicial no es crítica para la invención y no es el propósito de la presente memoria. Luego se codifica cada lectura como una posición en la secuencia de referencia y una lista de diferencias con dicha secuencia de referencia. Luego puede reconstruirse cada lectura a partir de la información codificada de alineación y la secuencia de referencia, mediante un software de descompresión adecuado configurado según la presente invención.

Preferiblemente, el software de alineación que procesa las lecturas y las alinea con la secuencia de referencia antes de proporcionarlas como entradas al software y aparato de compresión no tiene en cuenta determinados tipos de errores introducidos en las lecturas de secuencia, tales como, por ejemplo, errores de inserción o errores de deleción. Un error de inserción consiste en la inserción en una lectura de secuencia de uno o más símbolos adicionales que no se refieren a ningún ácido nucleico realmente presente. Un error de deleción consiste en la deleción de una lectura de secuencia de uno o más símbolos que representan ácidos nucleicos que están realmente presentes en la muestra secuenciada. Más precisamente, en el caso de un error de inserción o un error de deleción en una lectura de secuencia dada, el software de alineación considerará entonces los ácidos nucleicos erróneos resultantes como errores de sustitución, también denominados "apareamientos erróneos". Esta elección preferencial para la configuración del software de alineación permite una codificación posterior más rápida, proporcionando notablemente un mejor compromiso entre la velocidad y la razón compresión.

Para cada lectura, el software de alineación proporciona un registro de lectura al software y aparato de compresión. Cada registro de lectura contiene al menos la siguiente información: la posición de inicio absoluta de la lectura alineada con respecto a la secuencia de referencia, la longitud de la lectura, el tipo de alineación de la lectura, el número de apareamientos erróneos identificados en la lectura, y la posición relativa de dichos posibles apareamientos erróneos en la lectura (cuando sea apropiado).

El método de compresión según la presente invención se describirá ahora con referencia a la Figura 1. El método se realiza, por ejemplo, mediante un aparato 20 mostrado en la Figura 2. El aparato comprende al menos un procesador 22 y una memoria 24 acoplada operativamente al procesador 22 para formar un dispositivo informático. La memoria 24 puede almacenar un código de programa o software informático 26 que comprende instrucciones ejecutables por ordenador que, cuando se ejecutan por el procesador 22, hacen que el procesador 22 realice operaciones que comprenden los pasos del método de compresión según la invención.

El archivo inicial en el que las lecturas alineadas se almacenan como una lista de lecturas se almacenan, por ejemplo, en una memoria del aparato 20. Volviendo a la Figura 1, el método comprende preferiblemente un paso inicial 2 en donde la lista inicial de lecturas alineadas se divide en bloques de lecturas. Normalmente, la lista de lecturas alineadas se divide en bloques de 50.000 lecturas, este valor específico no se interpreta como limitativo del alcance de la presente invención que puede aplicarse de la misma manera con otros valores. Preferiblemente, los bloques de lecturas tienen el mismo tamaño de bloque. Cada bloque de lecturas comienza con un encabezado que contiene información necesaria para decodificar el bloque, tal como, por ejemplo, el tamaño en bytes del contenido del bloque, y/o un identificador del bloque o su contenido y/o el número de lecturas contenidas en el bloque. Esto permite el soporte para la concatenación del archivo comprimido, así como también las capacidades de flujo en continuo (conteniendo cada bloque de lecturas toda la información necesaria para decodificar las lecturas del bloque). Además, dado que el método de compresión puede realizarse en bloque después de bloque, esto también permite el procesamiento múltiple en los bloques de lecturas, permitiendo de ese modo la paralelización y alguna ganancia resultante en el tiempo de procesamiento. Si todas las lecturas de un bloque dado tienen la misma longitud, la longitud de lectura también se almacena en el encabezado, de lo contrario, una lista de cada longitud de lectura se almacena explícitamente durante el método de compresión.

Cada registro de lectura contiene información sobre el tipo de alineación de la lectura. Normalmente, pueden identificarse dos tipos principales de alineación: alineación perfecta y alineación imperfecta, más un tipo adicional correspondiente a una lectura "no correlacionada". La "alineación imperfecta" significa que la lectura contiene al menos un apareamiento erróneo distinto de un N, mientras que al menos una parte de la lectura coincide con una parte de la secuencia de referencia (según esta definición, una lectura imperfectamente correlacionada puede contener una o más N, siempre que también contenga uno o más apareamientos erróneos). En una realización a modo de ejemplo, cada registro de lectura comienza con los siguientes indicadores de bits, teniendo cada indicador de bits un valor entre dos valores posibles:

- un primer indicador de bits indicativo de una orientación directa o inversa frente a la secuencia de referencia,
- un segundo indicador de bits indicativo de una alineación perfecta o no,
- un tercer indicador de bits indicativo de si la lectura contiene al menos un N,
- un cuarto indicador de bits indicativo de si la información de posición está codificada en 16 bits o 32 bits.

Los siguientes pasos 4-12 se realizan en bloque de lecturas después del bloque de lecturas y se leen después de leer dentro de los bloques.

El método comprende un siguiente paso 4 de determinación, para cada lectura alineada, si dicha lectura está perfecta o imperfectamente correlacionada con la secuencia de referencia, o si dicha lectura no está correlacionada con la secuencia de referencia. Este paso de determinación 4 comprende, para cada lectura imperfectamente correlacionada, comparar 4A el número de apareamientos erróneos entre dicha lectura y la secuencia de referencia con un valor umbral. En una realización preferida, aunque no debe interpretarse como limitativo del alcance de la presente invención, dicho valor umbral es igual a 31. Este valor específico se ha elegido de manera intencionada para proporcionar el mejor compromiso posible para almacenar el número de apareamientos erróneos de una manera suficientemente compacta, tal como se entenderá mejor más adelante con respecto a el paso 12. De hecho, se ha observado estadísticamente que en una gran mayoría de los casos, las lecturas imperfectamente correlacionadas tienen menos de 31 apareamientos erróneos. El principio que se encuentra detrás de esa elección consiste en la codificación de la manera más compacta los casos más frecuentes, deja tener algunos muy pocos casos degradados. Si se determina que una lectura se correlaciona imperfectamente con un número de apareamientos erróneos inferiores al valor umbral, el paso de determinación 4 comprende una determinación adicional en cuanto a si la lectura se correlaciona global o localmente con la secuencia de referencia. Una "lectura globalmente correlacionada" es una lectura imperfectamente correlacionada cuya secuencia completa, que comprende el comienzo y el final de la lectura, se correlaciona imperfectamente con la secuencia de referencia. Una "lectura localmente correlacionada" es una

lectura imperfectamente correlacionada que contiene un segmento de nucleótidos o bases que se correlaciona imperfectamente con la secuencia de referencia. Dicho segmento de nucleótidos o bases corresponde así a una parte de la lectura inicial.

Preferiblemente, el método comprende además un paso 6 de determinar, para cada lectura alineada, si dicha lectura comprende al menos un N, es decir, si dicha lectura comprende al menos un apareamiento erróneo correspondiente a un caso en el que la máquina de secuenciación no pudo llamar a ninguna base o nucleótido. El método comprende entonces, para cada lectura que comprende al menos un N, un paso 8 de determinar el número de tales apareamientos erróneos N y un paso 10 de comparar dicho número de apareamientos erróneos N con un valor umbral de referencia. En una realización preferida, aunque no debe interpretarse como limitativo del alcance de la presente invención, dicho valor umbral es igual a 31.

Cualquiera que sea el resultado del paso de determinación 4, el método comprende un siguiente paso 12 de codificar las lecturas según dicha determinación al menos. Más precisamente, las lecturas que se determina que están perfectamente correlacionadas con la secuencia de referencia, si no comprenden ningún N o tiene un número de N menor que el valor umbral de referencia, se codifican según un primer proceso de codificación. Las lecturas que se determinan que están no correlacionadas o las lecturas que se determina que están perfectamente correlacionadas pero con un número de N mayor que el valor umbral de referencia se codifican según un segundo proceso de codificación en el que cada nucleótido o base se codifica individualmente, independientemente de si dicho nucleótido o base está alineado o no. Las lecturas que se determinan que se correlacionan imperfectamente se codifican según el segundo proceso de codificación o un tercer proceso de codificación. Más precisamente, las lecturas que se determinan que se correlacionan imperfectamente con un número de apareamientos erróneos mayor que el valor umbral se codifican según el segundo proceso de codificación. Si se determina que una lectura se correlaciona imperfectamente con un número de apareamientos erróneos inferior al valor umbral, si dicha lectura no comprende ningún N o tiene un número de N menor que el valor umbral de referencia, entonces dicha lectura se codifica según el tercer proceso de codificación. Si no, es decir, si la lectura tiene un número de N mayor que el valor umbral de referencia, entonces dicha lectura se codifica según el segundo proceso de codificación.

Ya sea que se haya determinado que una lectura dada está perfectamente correlacionada, imperfectamente correlacionada o no correlacionada, si dicha lectura comprende al menos un N pero tiene un número de N menor que el valor umbral de referencia, el paso de codificación 12 comprende codificar una lista de posiciones a lo largo de la secuencia de referencia, correspondiendo dichas posiciones a las posiciones de los N en la secuencia de referencia. Luego se almacena la lista de posiciones en una memoria de un dispositivo informático, implementando dicho dispositivo el método de compresión. Si una lectura comprende al menos un N pero tiene un número de N menor que el valor umbral de referencia, y debe codificarse según el segundo proceso de codificación, entonces cada nucleótido o base de la lectura se codifica individualmente en 2 bits.

Si una lectura comprende al menos un N pero con un número de N mayor que el valor umbral de referencia, entonces dicha lectura se codifica en cualquier caso según el segundo proceso de codificación, y cada nucleótido o base de la lectura se codifica individualmente en 4 bits. En este caso, el paso de codificación 12 no comprende codificar y almacenar una lista de posiciones de los N en la secuencia de referencia. De hecho, cada apareamiento erróneo N se codifica entonces directamente según el segundo proceso de codificación, de la misma manera que los otros nucleótidos o bases de la lectura.

El primer y tercer procesos de codificación comprenden conjuntos distintos de descriptores. Cada conjunto de descriptores representa unívocamente las lecturas asociadas al proceso de codificación correspondiente, siendo cada uno del primer y tercer procesos de codificación un proceso de codificación por entropía de información reducida. Más precisamente, el tercer proceso de codificación comprende un primer subproceso de codificación y un segundo subproceso de codificación. Las lecturas imperfectamente correlacionadas que se determinan que están globalmente correlacionadas durante el paso 4 se codifican según el primer subproceso de codificación. Las lecturas imperfectamente correlacionadas que se determinan que están localmente correlacionadas durante el paso 4 se codifican según el segundo subproceso de codificación. El primer y segundo subprocesos de codificación comprenden conjuntos distintos de descriptores, representando cada conjunto de descriptores unívocamente las lecturas asociadas al subproceso de codificación correspondiente.

La información de alineación codificada para cada lectura, y que permite la reconstrucción de la secuencia de lectura completa durante la descompresión de los datos, depende entonces del proceso o subproceso de codificación correspondiente usado para dicha lectura. Por ejemplo, los descriptores utilizados para el primer proceso de codificación pueden ser:

- la posición de inicio absoluta de la lectura perfectamente correlacionada con respecto a la secuencia de referencia (codificada en 16 o 32 bits), y

- la longitud de la lectura (codificada con la codificación diferencial con respecto a la longitud de la lectura previa, con código de longitud variable que varía de desde 2 bits hasta 34 bits).

Los descriptores utilizados para el primer subproceso de codificación pueden ser:

- la posición de inicio absoluta de la lectura imperfectamente correlacionada con respecto a la secuencia de referencia (codificada en 16 o 32 bits),
- la longitud de la lectura (codificada con la codificación diferencial con respecto a la longitud de la lectura previa, con código de longitud variable que varía de desde 2 bits hasta 34 bits), y
- una lista de las apareamientos erróneos de la lectura.

Los descriptores utilizados para el segundo subproceso de codificación pueden ser:

- la posición de inicio absoluta de la lectura imperfectamente correlacionada con respecto a la secuencia de referencia también denominada posición de inicio de alineación (codificada en 16 o 32 bits),
- la longitud de la lectura (codificada con la codificación diferencial con respecto a la longitud de la lectura previa, con código de longitud variable que varía de desde 2 bits hasta 34 bits),
- una lista de las apareamientos erróneos de la lectura, y
- la longitud de las partes recortadas de la lectura que no forman parte de la alineación (codificadas en 8 bits para cada parte recortada).

Preferiblemente, la lista de apareamientos erróneos que se codifica en el primer y segundo subprocesos comprende un encabezado (indicador de bits, codificado en 1 byte). Los cinco primeros bits del byte se usan para codificar el número de apareamientos erróneos contenidos en la lectura (en la realización preferida en donde el valor umbral es igual a 31, dicho número está dentro del intervalo [0-31]). Luego se usa un bit para codificar si la lectura imperfectamente correlacionada se correlaciona global o localmente. Otro bit se usa para codificar si el modo de 2 bits se activa o no para el segundo proceso de codificación. El último bit se usa para codificar si el modo de 4 bits se activa o no para el segundo proceso de codificación. Preferiblemente, para cada lectura codificada según el segundo subproceso de codificación durante el paso de codificación 12, las partes recortadas de dicha lectura (es decir, aquellas partes que no forman parte de la alineación local) se concatenan, y cada nucleótido o base de dichas partes recortadas se codifica individualmente. En una realización preferida, cada nucleótido o base de tales partes cortadas de la lectura se codifica individualmente en 2 bits.

En una realización preferida, cada apareamiento erróneo codificado en la lista de apareamientos erróneos de una lectura imperfectamente correlacionada (es decir, codificada según el primer o segundo subproceso de codificación) está codificada en 1 byte. Más precisamente, cada apareamiento erróneo de una lectura imperfectamente correlacionada que va a codificarse según el primer o segundo subproceso de codificación puede codificarse de la siguiente manera:

- los dos primeros bits del byte se usan para codificar el nucleótido o base alternativa presente en la lectura en lugar del nucleótido o base de referencia correspondiente en la secuencia de referencia,
- los seis últimos bits se utilizan para codificar la posición del apareamiento erróneo en la secuencia de referencia, calculándose dicha posición como un desplazamiento del apareamiento erróneo previo de la lectura (posición relativa del apareamiento erróneo, excepto por el primer apareamiento erróneo de la lectura para la que se codifica la posición absoluta). El intervalo de este desplazamiento, que se codifica en 6 bits, es por tanto de [0-63].

La Figura 3 proporciona un ejemplo de la codificación de los apareamientos erróneos de una lectura según el primer subproceso de codificación. La lectura es una lectura imperfectamente correlacionada, que se correlaciona globalmente con la secuencia de referencia. La lectura tiene dos apareamientos erróneos:

- un primer apareamiento erróneo, ubicado en la 12ª posición en la lectura, que consiste en una sustitución de un nucleótido A en la secuencia de referencia por un nucleótido T en la lectura, y
- un segundo apareamiento erróneo, ubicado en la 21ª posición en la lectura, que consiste en una sustitución de un nucleótido C en la secuencia de referencia por un nucleótido G en la lectura.

La lista de apareamientos erróneos de la lectura se codifica entonces como:

- <12, T>, el valor “12” correspondiente a la posición absoluta del primer apareamiento erróneo en la lectura, y
- <9, G>, el valor “9” correspondiente a la posición relativa del segundo apareamiento erróneo en la lectura, es decir, el desplazamiento entre el segundo apareamiento erróneo y el primer apareamiento erróneo.

<12, T>, por ejemplo, puede convertirse en el valor "51" (codificado en 1 byte), y <9, G> puede convertirse en el valor "38" (codificado en 1 byte). Una codificación de bytes de este tipo se obtiene con:

posición de desplazamiento x 4 + valor de nucleótido (con A=0, C=1, G=2, T=3)

Preferiblemente, para cada lectura imperfectamente correlacionada que va a codificarse según el primer o segundo subproceso de codificación, si el desplazamiento calculado entre un apareamiento erróneo dado de la lectura y el apareamiento erróneo previo es mayor que un valor codificable máximo, entonces se inserta al menos un apareamiento erróneo "falso" entre dichos dos apareamientos erróneos hasta que cada desplazamiento entre cada uno de dichos apareamientos erróneos y el al menos un apareamiento erróneo "falso" es menor que dicho valor codificable máximo. Un apareamiento erróneo "falso" se define como un apareamiento erróneo para el cual los bits del byte usado para codificar el apareamiento erróneo codifican un nucleótido o base que es igual al nucleótido o base de referencia correspondiente en la secuencia de referencia. En una realización preferida, aunque no debe interpretarse como limitativo del alcance de la presente invención, el valor codificable máximo es igual a 63, correspondiente al valor máximo que puede codificarse en 6 bits.

La Figura 4 proporciona un ejemplo de la codificación de los apareamientos erróneos de una lectura según el primer subproceso de codificación, en un caso en el que debe insertarse un apareamiento erróneo "falso". La lectura es una lectura imperfectamente correlacionada, que se correlaciona globalmente con la secuencia de referencia. La lectura tiene dos apareamientos erróneos:

- un primer apareamiento erróneo, ubicado en la 22ª posición en la lectura, que consiste en una sustitución de un nucleótido A en la secuencia de referencia por un nucleótido T en la lectura, y

- un segundo apareamiento erróneo, ubicado en la 134ª posición en la lectura, que consiste en una sustitución de un nucleótido C en la secuencia de referencia por un nucleótido G en la lectura.

El desplazamiento de posición entre el segundo y el primera apareamientos erróneos es de 112, que es mayor que el valor codificable máximo de 63. Por tanto, tiene que insertarse un apareamiento erróneo "falso" entre los dos apareamientos erróneos, de modo que cada desplazamiento entre cada uno de los apareamientos erróneos y el apareamiento erróneo "falso" es menor que dicho valor codificable máximo. Un apareamiento erróneo "falso" con un nucleótido T (correspondiente a un nucleótido T "real" en la secuencia de referencia) se inserta, por ejemplo, en la 85ª posición en la lectura. El desplazamiento de posición calculado entre el apareamiento erróneo "falso" y el primer apareamiento erróneo es de 63, que corresponde al valor codificable máximo. El desplazamiento de posición calculado entre el segundo apareamiento erróneo y el apareamiento erróneo "falso" es de 49, que es menor de 63.

La lista de apareamientos erróneos de la lectura se codifica entonces como:

- <22, T>, el valor "22" correspondiente a la posición absoluta del primer apareamiento erróneo en la lectura,

- <63, T>, el valor "63" correspondiente a la posición relativa del apareamiento erróneo "falso" en la lectura, es decir, el desplazamiento entre el apareamiento erróneo "falso" y el primer apareamiento erróneo, y

- <49, G>, el valor "49" correspondiente a la posición relativa del segundo apareamiento erróneo en la lectura, es decir, el desplazamiento entre el segundo apareamiento erróneo y el apareamiento erróneo "falso".

<22, T> puede convertirse, por ejemplo, en el valor "91" (codificado en 1 byte), <63, T> puede convertirse en el valor "255" (codificado en 1 byte), y <49, G> puede convertirse en el valor "198" (codificado en 1 byte). Una codificación de bytes de este tipo se obtiene con:

posición de desplazamiento x 4 + valor de nucleótido (con A=0, C=1, G=2, T=3)

El método comprende un paso final 14 de proporcionar un archivo comprimido que comprende una lista de lecturas codificadas. Las lecturas codificadas se almacenan en el archivo comprimido en el mismo orden que las lecturas almacenadas en el archivo inicial sin comprimir. Luego puede reconstruirse cada lectura a partir de la información codificada de alineación y la secuencia de referencia, mediante un software y/o método de descompresión adecuado configurado según la presente invención.

Aunque se describe con referencia a una arquitectura a modo de ejemplo de un dispositivo informático 20 (mostrado en la Figura 2 con fines ilustrativos), las técnicas de la invención descritas en el presente documento pueden implementarse en hardware, software, firmware o cualquier combinación de los mismos. Cuando se implementa en software, el código de programa informático puede almacenarse en un medio informático y ejecutarse mediante una unidad de procesamiento de hardware que comprende uno o más procesadores, como es el caso con el dispositivo 20 de la Figura 2. Debe entenderse que el término "procesador", tal como se usa en el presente documento, pretende incluir uno o más dispositivos de procesamiento, incluyendo un procesador de señales, un microprocesador, un

microcontrolador, un circuito integrado de aplicación específica (ASIC), una matriz de puertas programables en campo (FPGA) u otro tipo de conjuntos de circuitos de procesamiento, así como partes o combinaciones de tales elementos de conjuntos de circuitos. Además, el término "memoria" tal como se usa en el presente documento pretende incluir memoria electrónica asociada con un procesador, tal como memoria de acceso aleatorio (RAM), memoria de solo lectura (ROM) u otros tipos de memoria, en cualquier combinación.

En consecuencia, las instrucciones o código de software para realizar las metodologías y protocolos descritos en el presente documento pueden almacenarse en uno o más de los dispositivos de memoria asociados, por ejemplo, ROM, memoria fija o extraíble, y, cuando están listos para utilizarse, se cargan en la RAM y son ejecutados por el procesador.

Las técnicas de esta descripción pueden implementarse en una amplia variedad de dispositivos o aparatos, incluyendo, por ejemplo, teléfonos móviles, ordenadores, servidores, tabletas y dispositivos similares.

Ejemplos estadísticos y numéricos del método de compresión según la invención

El siguiente ejemplo comparativo se ha realizado en un archivo de datos no comprimido que contenía 48 millones de lecturas o secuencias de nucleótidos:

- tamaño del archivo de datos no comprimido: 35.770 MB (MegaByte)
- el tamaño del archivo que se ha comprimido con el software gzip 6.649 MB
- tamaño del archivo que se ha comprimido con el software SPRING no basado en referencia: 1.402 MB
- tamaño del archivo que se ha comprimido con el método de compresión basado en referencia según la presente invención: 1.179 MB
- tiempo de compresión con el software SPRING no basado en referencia; 1.722 s
- tiempo de compresión con el método de compresión basado en referencia según la presente invención: 181 s
- tamaño promedio en bit/nucleótido del archivo de datos no comprimido (codificación ASCII): 8 bit/nucleótido
- tamaño promedio en bit/nucleótido del archivo que se ha comprimido con una codificación adaptada a 4 caracteres posibles A, T, C, G: 2 bit/nucleótido
- tamaño promedio en bit/nucleótido del archivo que se ha comprimido con el método de compresión basado en referencia según la presente invención: 0,33 bit/nucleótido

Los ejemplos numéricos indicados anteriormente ilustran que la presente invención permite una compresión y descompresión rápidas, proporcionando al mismo tiempo una alta razón de compresión.

REIVINDICACIONES

1. Un método para comprimir datos de secuencia genómica, comprendiendo el método:

obtener, mediante uno o más ordenadores, un registro de lectura;
determinar, mediante uno o más ordenadores, si el registro de lectura corresponde a una lectura que se correlaciona perfectamente con una secuencia de referencia o se correlaciona imperfectamente con la secuencia de referencia;
basándose en la determinación, mediante el uno o más ordenadores, de que el registro de lectura corresponde a una lectura que se correlaciona imperfectamente con la secuencia de referencia, determinar, mediante el uno o más ordenadores, si un número de apareamientos erróneos de la lectura imperfectamente correlacionada satisface un número umbral predeterminado de apareamientos erróneos; y
basándose en la determinación de que el número de apareamientos erróneos satisface el número umbral predeterminado de apareamientos erróneos, (i) obtener, mediante el uno o más ordenadores, un desplazamiento de un apareamiento erróneo previo que es menor que un valor de desplazamiento que puede codificarse máximo y (ii) codificar, mediante el uno o más ordenadores, cada apareamiento erróneo de la lectura imperfectamente correlacionada y el desplazamiento del apareamiento erróneo previo de la lectura en un registro que tiene un tamaño de 1 byte.

2. El método de la reivindicación 1, en donde determinar, mediante el uno o más ordenadores, si un número de apareamientos erróneos de la lectura imperfectamente correlacionada satisface un número umbral predeterminado de apareamientos erróneos comprende:
determinar, mediante el uno o más ordenadores, si el número de apareamientos erróneos de la lectura imperfectamente correlacionada es mayor que el número umbral predeterminado de apareamientos erróneos.

3. El método de la reivindicación 1, en donde cada registro de lectura comprende:

datos que indican una posición de inicio absoluta de una lectura alineada con respecto a la secuencia de referencia;
datos que indican una longitud de la lectura;
datos que indican si la lectura está perfectamente correlacionada o imperfectamente correlacionada;
datos que indican un número de apareamientos erróneos identificados en la lectura; y
datos que indican una posición relativa de cada uno de dichos apareamientos erróneos en la lectura.

4. El método de la reivindicación 1, en donde codificar cada apareamiento erróneo de la lectura imperfectamente correlacionada en un registro que tiene un tamaño de 1 byte comprende para cada apareamiento erróneo particular:

codificar, mediante el uno o más ordenadores, los dos primeros bits del byte para incluir datos que representan un nucleótido o base alternativa presente en la lectura en lugar de un nucleótido o base de referencia correspondiente en la secuencia de referencia; y
codificar, mediante el uno o más ordenadores, los seis bits restantes del byte para incluir datos que representan el desplazamiento.

5. El método de la reivindicación 4, comprendiendo el método además:

determinar, mediante uno o más ordenadores, si el desplazamiento es mayor que un valor codificable máximo; y
basándose en la determinación de que el desplazamiento es mayor que el valor codificable máximo, insertar, mediante uno o más ordenadores, al menos un falso apareamiento erróneo entre el apareamiento erróneo particular y el apareamiento erróneo previo.

6. El método de la reivindicación 1, comprendiendo el método además:

basándose en la determinación de que el número de apareamientos erróneos no satisface el número umbral predeterminado de apareamientos erróneos, codificar, mediante uno o más ordenadores, una lista de posiciones de la secuencia de referencia correspondiente a una posición de cada uno de los apareamientos erróneos con respecto a la secuencia de referencia usando un proceso de codificación por entropía de información reducida.

7. El método de la reivindicación 1, comprendiendo el método además:
basándose en la determinación de que el registro de lectura corresponde a una lectura que se correlaciona perfectamente con la secuencia de referencia, codificar, mediante uno o más ordenadores, al menos una parte del registro de lectura usando codificación por entropía de información reducida.

8. El método de la reivindicación 1, en donde el uno o más ordenadores comprenden uno o más procesadores de hardware.

9. El método de la reivindicación 8, en donde el uno o más procesadores de hardware comprenden una o más matrices de puertas programables en campo (FPGA).

10. Un procesador de hardware que incluye un conjunto de circuitos de procesamiento de hardware que se configura para realizar una o más operaciones, comprendiendo la una o más operaciones un método según cualquiera de las reivindicaciones anteriores.

11. Un sistema para comprimir datos de secuencia genómica, comprendiendo el sistema:

uno o más ordenadores y uno o más dispositivos de almacenamiento que almacenan instrucciones que pueden funcionar, cuando se ejecutan mediante uno o más ordenadores, para hacer que uno o más ordenadores realicen las operaciones que comprenden un método según una cualquiera de las reivindicaciones 1 a 9.

12. Un dispositivo de almacenamiento legible por ordenador que tiene instrucciones almacenadas en el mismo, que, cuando se ejecutan mediante un aparato de procesamiento de datos, hacen que el aparato de procesamiento de datos realice operaciones para comprimir datos de secuencia genómica, comprendiendo las operaciones un método según una cualquiera de las reivindicaciones 1 a 9.

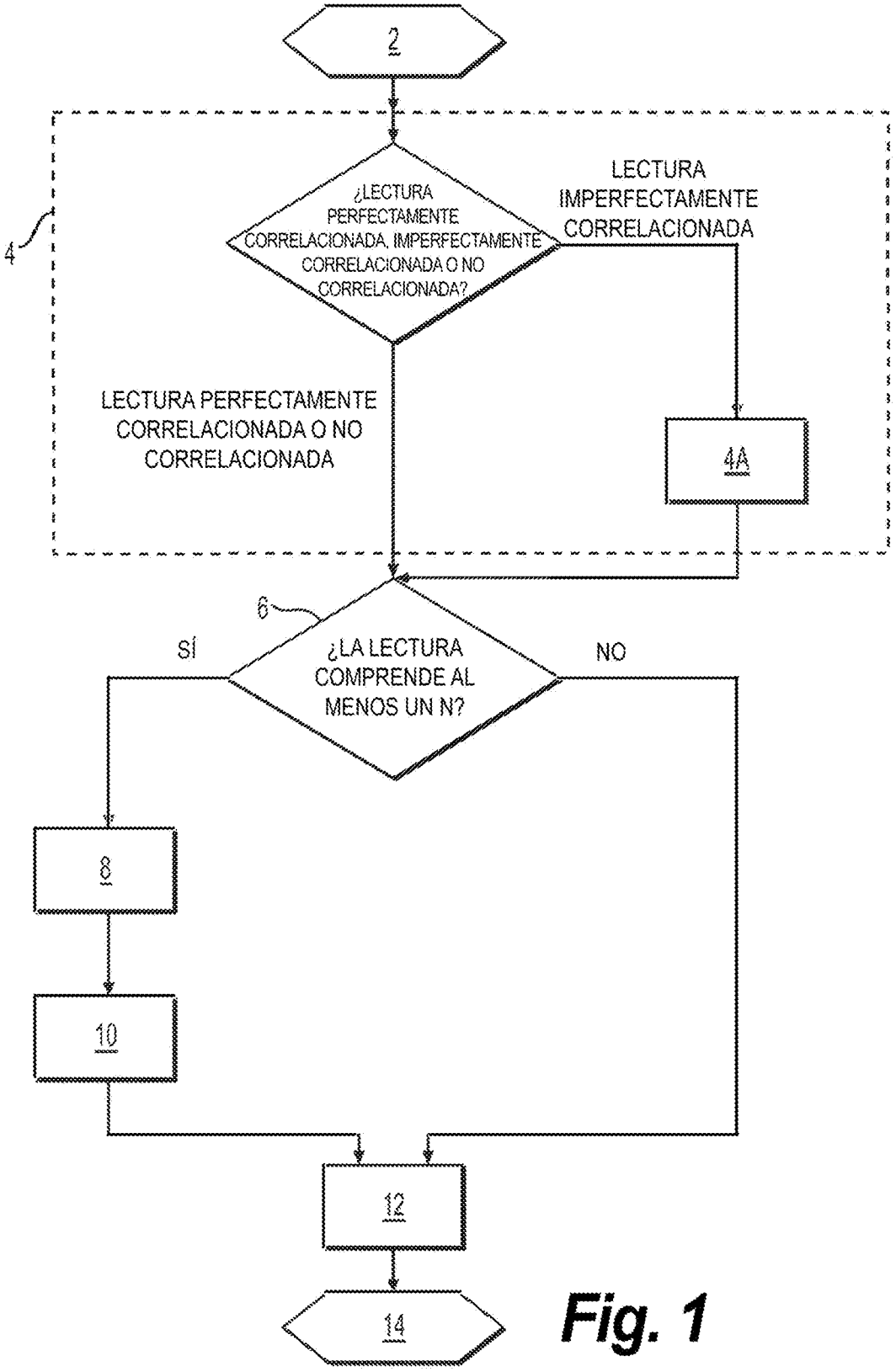
2
4
¿LECTURA PERFECTAMENTE CORRELACIONADA, IMPERFECTAMENTE CORRELACIONADA O NO CORRELACIONADA?
LECTURA IMPERFECTAMENTE CORRELACIONADA
LECTURA PERFECTAMENTE CORRELACIONADA O NO CORRELACIONADA
4A
6
¿LA LECTURA COMPRENDE AL MENOS UN N?
SÍ
NO
8
10
12
14

Fig. 1

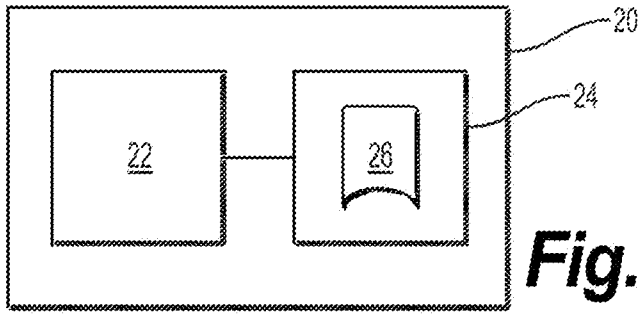

Fig. 2

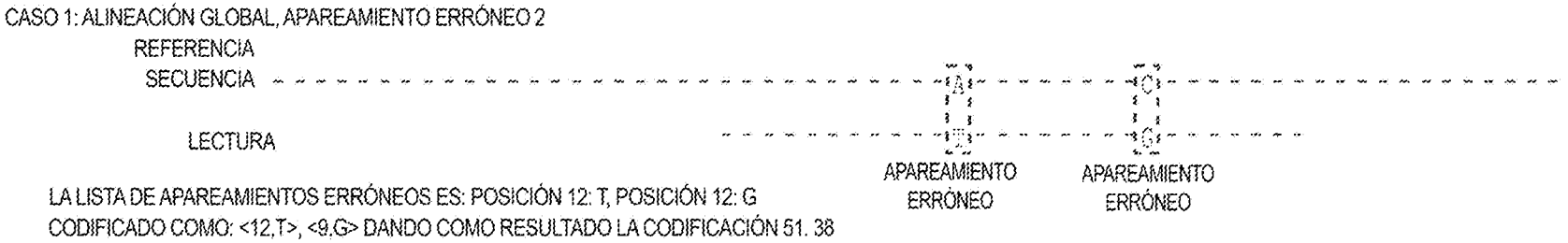

Fig. 3

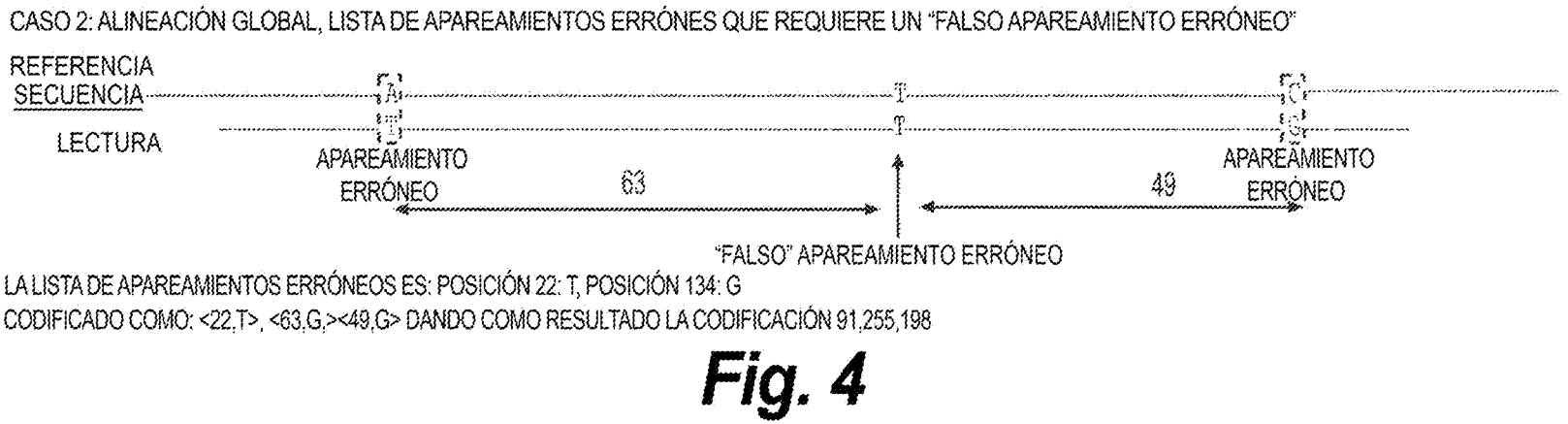

Fig. 4

ES 2 964 351 T3