(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 978 795**

(51) Int. Cl.:

G16B 35/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **29.01.2014 PCT/US2014/013666**

(87) Fecha y número de publicación internacional: **07.08.2014 WO14120819**

(96) Fecha de presentación y número de la solicitud europea: **29.01.2014 E 14746677 (5)**

(97) Fecha y número de publicación de la concesión europea: **20.03.2024 EP 2951754**

(54) Título: **Métodos, sistemas y software para identificar biomoléculas con componentes que interactúan**

(30) Prioridad:

31.01.2013 US 201361759276 P
15.03.2013 US 201361799377 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
20.09.2024

(73) Titular/es:

CODEXIS, INC. (100.0%)
200 Penobscot Drive
Redwood City, CA 94063, US

(72) Inventor/es:

COPE, GREGORY ALLAN

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Métodos, sistemas y software para identificar biomoléculas con componentes que interactúan

ANTECEDENTES

La presente divulgación está relacionada con los campos de la biología molecular, la evolución molecular, la bioinformática y los sistemas digitales. Más específicamente, la divulgación se refiere a métodos para predecir computacionalmente la actividad de una biomolécula y/o guiar la evolución dirigida. También se proporcionan sistemas, incluyendo sistemas digitales, y software de sistema para llevar a cabo estos métodos. Los métodos de la presente divulgación tienen utilidad en la optimización de proteínas para uso industrial y terapéutico.

Desde hace tiempo se sabe que el diseño de proteínas es una tarea difícil, aunque sólo sea por la explosión combinatoria de posibles moléculas que constituyen el espacio de secuencias susceptible de búsqueda. El espacio de secuencias de las proteínas es inmenso e imposible de explorar exhaustivamente usando los métodos actualmente conocidos en la técnica. Debido a esta complejidad, se han usado muchos métodos aproximados para diseñar mejores proteínas; el principal de ellos es el método de la evolución dirigida. Hoy en día, la evolución dirigida de proteínas está dominada por varios formatos de cribado y recombinación de alto rendimiento, a menudo realizados iterativamente.

Paralelamente, se han propuesto varias técnicas computacionales para explorar el espacio de actividad de secuencia. Aunque cada técnica computacional presenta ventajas en ciertos contextos, sería muy deseable disponer de nuevas maneras de buscar eficientemente en el espacio de secuencias para identificar proteínas funcionales.

La US 2005/0084907 se refiere a métodos para buscar rápida y eficazmente el espacio de datos relacionados con la biología.

SUMARIO

La invención proporciona un método implementado por ordenador para identificar moléculas biológicas con actividad deseada mejorada, el método comprendiendo:

(a) recibir datos de secuencia y actividad de una pluralidad de moléculas biológicas;
(b) preparar un modelo base a partir de los datos de secuencia y actividad, en donde el modelo base predice la actividad en función de la presencia o ausencia de subunidades de la secuencia y el modelo base no incluye términos de interacción de un conjunto definido de términos de interacción;
(b1) establecer un modelo de secuencia actual y un modelo de mejor secuencia para el modelo base;
(c1) crear un nuevo modelo de secuencia añadiendo al modelo de secuencia actual un término de interacción todavía no añadido del conjunto definido de términos de interacción;
(c2) evaluar la potencia predictiva del nuevo modelo de secuencia, y si la potencia predictiva del nuevo modelo de secuencia es mayor que el del mejor modelo de secuencia, establecer el nuevo modelo de secuencia como el mejor modelo de secuencia;
(c3) si hay algún término de interacción en el conjunto definido de términos de interacción que no se haya añadido al modelo de secuencia actual, repetir (c1)-(c3);
(d) si se estableció un nuevo modelo como mejor modelo de secuencia en (c2), establecer el modelo de secuencia actual como mejor modelo de secuencia y repetir (c1)-(d);
(e) establecer el mejor modelo de secuencia como modelo final; y
(f) usar el modelo final para guiar la evolución dirigida y seleccionar una o más secuencias con la actividad deseada mejorada,
caracterizado porque el término de interacción se selecciona aleatoriamente en el paso (c1).

La invención proporciona además un método implementado por ordenador para identificar moléculas biológicas con actividad deseada mejorada, el método comprendiendo:

(a) recibir datos de secuencia y actividad para una pluralidad de moléculas biológicas;
(b) preparar un modelo base a partir de los datos de secuencia y actividad, en donde el modelo base predice la actividad en función de la presencia o ausencia de subunidades de la secuencia y el modelo base incluye todos los términos de interacción de un conjunto definido de términos de interacción;
(b1) establecer un modelo de secuencia actual y un modelo de secuencia mejor en el modelo base;
(c1) crear un nuevo modelo de secuencia sustrayendo del modelo de secuencia actual un término de interacción todavía no sustraído del conjunto definido de términos de interacción;
(c2) evaluar la potencia predictiva del nuevo modelo de secuencia, y si la potencia predictiva del nuevo modelo de secuencia es mayor que el del mejor modelo de secuencia establecer el nuevo modelo de secuencia como el mejor modelo de secuencia;
(c3) si hay algún término de interacción en el conjunto definido de términos de interacción que no se haya sustraído del modelo de secuencia actual, repetir (c1)-(c3);

(d) si se estableció un nuevo modelo como mejor modelo de secuencia en (c2), establecer el modelo de secuencia actual como mejor modelo de secuencia y repetir (c1)-(d);
(e) establecer el mejor modelo de secuencia como modelo final; y
(f) usar el modelo final para guiar la evolución dirigida y seleccionar una o más secuencias con la actividad deseada mejorada,
caracterizado porque el término de interacción se selecciona aleatoriamente en el paso (c1).

La invención proporciona además un producto de programa informático que comprende uno o más medios de almacenamiento no transitorios legibles por ordenador que tienen almacenadas en los mismos instrucciones ejecutables por ordenador que, cuando son ejecutadas por uno o más procesadores de un sistema informático, hacen que el sistema informático implemente el método de la invención.

La invención proporciona además un sistema informático, que comprende:

uno o más procesadores;
memoria del sistema; y
uno o más medios de almacenamiento legibles por ordenador que tienen almacenadas en los mismos instrucciones ejecutables por ordenador que, cuando son ejecutadas por el uno o más procesadores, hacen que el sistema informático implemente el método de la invención.

La presente divulgación presenta técnicas para generar y usar modelos de secuencia-actividad que emplean términos no lineales, en particular términos que tienen en cuenta las interacciones entre dos o más subunidades de una secuencia, como se define adicionalmente en las reivindicaciones. Los modelos de secuencia-actividad describen actividades, características o propiedades de moléculas biológicas como funciones de varias secuencias biológicas. Estos términos no lineales pueden ser términos de "producto cruzado" que implican la multiplicación de dos o más variables, cada una de las cuales representa la presencia (o ausencia) de las subunidades que participan en la interacción. Algunas realizaciones implican técnicas para seleccionar los términos no lineales que mejor describen la actividad de la secuencia. Tener en cuenta que a menudo hay muchos más términos de interacción no lineal posibles que verdaderas interacciones entre subunidades. Por lo tanto, para evitar el sobreajuste, típicamente sólo se considera un número limitado de términos no lineales y los empleados deben reflejar las interacciones que afectan apreciablemente a la actividad.

Un aspecto de la divulgación, que se define adicionalmente en las reivindicaciones, proporciona un método para preparar un modelo de secuencia-actividad que puede ayudar en la identificación de moléculas biológicas con actividad deseada mejorada, el método comprendiendo: (a) recibir datos de secuencia y actividad para una pluralidad de moléculas biológicas; (b) preparar un modelo base a partir de los datos de secuencia y actividad, en donde el modelo base predice la actividad como una función de la presencia o ausencia de subunidades de la secuencia; (c) preparar por lo menos un nuevo modelo añadiendo o sustrayendo por lo menos un nuevo término de interacción a o del modelo base, en donde el nuevo término de interacción representa la interacción entre dos o más subunidades que interactúan; (d) determinar la capacidad del por lo menos un nuevo modelo para predecir la actividad en función de la presencia o ausencia de las subunidades; y (e) determinar si añadir o sustraer el nuevo término de interacción al o del modelo base basándose en la capacidad del por lo menos un nuevo modelo para predecir la actividad según lo determinado en (d) y con un sesgo en contra de añadir el nuevo término de interacción. El modelo derivado puede usarse luego en varias aplicaciones, como en la evolución dirigida de bibliotecas de proteínas para identificar proteínas con las actividades y propiedades biológicas deseadas.

En donde el método determina que el nuevo término de interacción debería añadirse al modelo base para producir un modelo actualizado, el método incluye además pasos adicionales para buscar términos de interacción adicionales que puedan mejorar aún más el modelo actualizado. Específicamente, el método incluye: (f) repetir (c) usando el modelo actualizado en lugar del modelo base y añadir o sustraer un término de interacción diferente del añadido/sustraído en (c); y (g) repetir (d) y (e) usando el modelo actualizado en lugar del modelo base. En algunas realizaciones, el método incluye además (h) repetir (f) y (g) usando otro modelo actualizado. En varias realizaciones, la secuencia puede ser un genoma completo, un cromosoma completo, un segmento cromosómico, una colección de secuencias génicas para genes que interactúan, un gen, una secuencia de ácido nucleico, una proteína, un polisacárido, etc. En una o más realizaciones, las subunidades de las secuencias pueden ser cromosomas, segmentos cromosómicos, haplotipos, genes, nucleótidos, codones, mutaciones, aminoácidos, carbohidratos (mono, di, tri u oligoméricos), etc.

En una o más implementaciones consistentes con las realizaciones anteriores, se proporciona un método para identificar residuos de aminoácidos a modificar en una biblioteca de variantes de proteínas. En estas realizaciones, una pluralidad de moléculas biológicas constituye un conjunto de entrenamiento de una biblioteca de variantes de proteínas. La biblioteca de variantes de proteínas puede incluir proteínas de varias fuentes. En un ejemplo, los miembros incluyen proteínas de origen natural como las codificadas por miembros de una única familia de genes. En otro ejemplo, las secuencias incluyen proteínas obtenidas usando un mecanismo de generación de diversidad basado en la recombinación. Por ejemplo, la recombinación mediada por fragmentación del ADN, la recombinación

mediada por oligonucleótidos sintéticos o una combinación de las mismas puede realizarse sobre ácidos nucleicos que codifican la totalidad o parte de una o más proteínas parentales naturales para este propósito. En otro ejemplo más, los miembros se obtienen implementando un protocolo de diseño de experimentos (DOE) para identificar las secuencias variadas sistemáticamente.

En algunas realizaciones, por lo menos un término de interacción es un término de producto cruzado que contiene un producto de una variable que representa la presencia de un residuo que interactúa y otra variable que representa la presencia de otro residuo que interactúa. La forma del modelo secuencia-actividad puede ser una suma de por lo menos un término de producto cruzado y uno o más términos lineales, representando cada uno de los términos lineales el efecto de un residuo variable en un conjunto de entrenamiento de una biblioteca de variantes de proteínas. El por lo menos un término de producto cruzado puede seleccionarse de un grupo de posibles términos de productos cruzados mediante varias técnicas, incluyendo la suma o resta de términos por pasos sin sustitución.

En una o más realizaciones, un modelo que incluye términos de producto cruzado se ajusta a los datos dados usando técnicas de regresión bayesianas, en las que se usa el conocimiento previo para determinar las distribuciones de probabilidad posteriores del modelo.

En una o más realizaciones, se crean dos o más modelos nuevos, cada uno de los cuales incluye por lo menos un término de interacción diferente. En tales realizaciones, el método comprende además la preparación de un modelo de conjunto basado en los dos o más modelos nuevos. El modelo de conjunto incluye términos de interacción de los dos o más modelos nuevos. El modelo de conjunto pondera los términos de interacción de acuerdo con las capacidades de los dos o más modelos nuevos para predecir la actividad de interés.

El modelo secuencia-actividad puede producirse a partir del conjunto de entrenamiento mediante muchas técnicas diferentes. En algunas realizaciones, el modelo es un modelo de regresión, como un modelo de mínimos cuadrados parciales, un modelo de regresión bayesiano o un modelo de regresión de componentes principales. En otra realización, el modelo es una red neuronal.

El uso del modelo secuencia-actividad para identificar residuos de fijación o variación puede implicar cualquiera de las muchas técnicas analíticas posibles diferentes. En algunos casos, se usa una "secuencia de referencia" para definir las variaciones. Dicha secuencia puede ser una que el modelo ha predicho que tiene el valor más alto (o uno de los valores más altos) de la actividad deseada. En otro caso, la secuencia de referencia puede ser la de un miembro de la biblioteca original de variantes de proteínas. A partir de la secuencia de referencia, el método puede seleccionar subsecuencias para efectuar las variaciones. Adicional o alternativamente, el modelo secuencia-actividad clasifica las posiciones de los residuos (o residuos específicos en ciertas posiciones) por orden de impacto sobre la actividad deseada.

Un objetivo del método puede ser generar una nueva biblioteca de variantes de proteínas. Como parte de este proceso, el método puede identificar secuencias que se usarán para generar esta nueva biblioteca. Tales secuencias incluyen variaciones de los residuos identificados en (e), (g) o (h) o son precursores usados para introducir posteriormente dichas variaciones. Las secuencias pueden modificarse realizando mutagénesis o un mecanismo de generación de diversidad basado en la recombinación para generar la nueva biblioteca de variantes de proteínas. Esto puede formar parte de un procedimiento de evolución dirigida. La nueva biblioteca también puede usarse para desarrollar un nuevo modelo de secuencia-actividad. La nueva biblioteca de variantes proteínicas se analiza para evaluar los efectos sobre una actividad concreta, como la estabilidad, la actividad catalítica, la actividad terapéutica, la resistencia a un patógeno o toxina, la toxicidad, etc.

En algunas realizaciones, el método implica seleccionar uno o más miembros de la nueva biblioteca de variantes de proteínas para su producción. Uno o más de ellos pueden sintetizarse y/o expresarse luego en un sistema de expresión. En una realización específica, el método continúa de la siguiente manera: (i) proporcionar un sistema de expresión a partir del cual pueda expresarse un miembro seleccionado de la nueva biblioteca de variantes de proteínas; y (ii) expresar el miembro seleccionado de la nueva biblioteca de variantes de proteínas.

En algunas realizaciones, en lugar de usar secuencias de aminoácidos, los métodos emplean secuencias de nucleótidos para generar los modelos y predecir la actividad. Las variaciones en grupos de nucleótidos, por ejemplo, codones, afectan a la actividad de los péptidos codificados por las secuencias de nucleótidos. En algunas realizaciones, el modelo puede proporcionar un sesgo para los codones que se expresan preferentemente (en comparación con otros codones que codifican el mismo aminoácido) dependiendo del huésped empleado para expresar el péptido.

Otro aspecto de la divulgación (como se define en las reivindicaciones) se refiere a aparatos y productos de programas informáticos que incluyen medios legibles por máquina en los que se proporcionan instrucciones de programa y/o disposiciones de datos para implementar los métodos y sistemas de software descritos anteriormente. Con frecuencia, las instrucciones del programa se proporcionan como código para realizar ciertas operaciones del método. Los datos, si se emplean para implementar características de esta divulgación, pueden proporcionarse como

estructuras de datos, tablas de bases de datos, objetos de datos u otras disposiciones apropiadas de información especificada. Cualquiera de los métodos o sistemas descritos en la presente puede representarse, en su totalidad o en parte, como tales instrucciones de programa y/o datos proporcionados en cualquier medio legible por máquina adecuado.

Estas y otras características se describen con más detalle a continuación en la descripción detallada y junto con las siguientes figuras.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La Figura 1 ilustra un método general por pasos para preparar un modelo secuencia-actividad.

La Figura 2 es un diagrama de flujo que representa una secuencia de operaciones para generar una o más generaciones de bibliotecas de variantes de proteínas, en donde las operaciones usan un modelo de secuencia-actividad como uno de los obtenidos en la Figura 1 para guiar la generación de bibliotecas de variantes de proteínas. Las bibliotecas de variantes generadas pueden proporcionar datos de secuencia y actividad para preparar uno o más nuevos modelos de secuencia-actividad, formando un bucle de modelado-exploración de evolución guiada.

Las Figuras 3A-3H son gráficos que muestran ejemplos en los que se comparan las capacidades predictivas de ciertos modelos lineales y no lineales.

Las Figura 4A-4B ilustran diagramas de flujo de procesos que implementan métodos de adición y sustracción por pasos para preparar un modelo de secuencia-actividad. La Figura 4A ilustra un ejemplo específico de un método de adición por pasos para preparar un modelo; y la Figura 4B ilustra un ejemplo específico de un método de sustracción por pasos para preparar un modelo.

La Figura 5 ilustra un diagrama de flujo de un proceso que implementa la regresión bayesiana en la evolución dirigida de variantes de secuencia de acuerdo con una realización.

La Figura 6 ilustra un diagrama de flujo de un proceso que implementa la regresión por conjuntos en la evolución dirigida de variantes de secuencias de acuerdo con una realización.

La Figura 7 es un diagrama de flujo que representa un método de p-valor de Bootstrap para generar bibliotecas de variantes de proteínas de acuerdo con una realización.

La Figura 8 es un esquema de un dispositivo digital ejemplar.

DESCRIPCIÓN DETALLADA

I. DEFINICIONES

A menos que se definan de otro modo en la presente, todos los términos técnicos y científicos usados en la presente tienen el mismo significado que el entendido comúnmente por un experto en la técnica. Varios diccionarios científicos que incluyen los términos incluidos en la presente son bien conocidos y están a disposición de los expertos en la técnica. Cualquier método y material similar o equivalente a los descritos en la presente puede usarse en la puesta en práctica de las realizaciones divulgadas en la presente.

Los términos definidos inmediatamente a continuación se entienden mejor si se hace referencia a la memoria descriptiva en su conjunto. Las definiciones tienen el propósito de describir únicamente realizaciones particulares y ayudar a comprender los complejos conceptos descritos en esta memoria descriptiva. No se pretende que limiten el alcance completo de la divulgación. Específicamente, debe entenderse que esta divulgación no se limita a las secuencias, composiciones, algoritmos, sistemas, metodología, protocolos y reactivos particulares descritos, ya que éstos pueden variar, dependiendo del contexto en el que sean usados por los expertos en la técnica.

Como se usan en esta memoria descriptiva y en las reivindicaciones adjuntas, las formas singulares "un", "uno" y "el" incluyen referentes plurales a menos que el contenido y el contexto indiquen claramente lo contrario. Así, por ejemplo, la referencia a "un dispositivo" incluye una combinación de dos o más de tales dispositivos, y similares.

A menos que se indique lo contrario, se pretende que una conjunción "o" se use en su sentido correcto como operador lógico booleano, abarcando tanto la selección de características en la alternativa (A o B, donde la selección de A es mutuamente excluyente de B) como la selección de características en conjunción (A o B, donde se seleccionan tanto A como B). En algunos lugares del texto, se usa el término "y/o" con el mismo propósito, lo que no debe interpretarse como que "o" se usa con referencia a alternativas mutuamente excluyentes.

Una "biomolécula" o "molécula biológica" es una molécula que generalmente se encuentra en un organismo biológico. En algunas realizaciones, las moléculas biológicas comprenden macromoléculas biológicas poliméricas que tienen múltiples subunidades (es decir, "biopolímeros"). Las biomoléculas típicas incluyen, pero no se limitan a, moléculas que comparten algunas características estructurales con polímeros naturales como los ARN (formados a partir de subunidades de nucleótidos), ADN (formados a partir de subunidades de nucleótidos) y péptidos o polipéptidos (formados a partir de subunidades de aminoácidos), incluyendo, por ejemplo, ARN, análogos de ARN, ADN, análogos de ADN, polipéptidos, análogos de polipéptidos, ácidos nucleicos peptídicos (ANP), combinaciones de

ARN y ADN (por ejemplo, quimeraplastos), o similares. No se pretende que las biomoléculas se limiten a ninguna molécula en particular, ya que cualquier molécula biológica adecuada encuentra uso en la presente invención, incluyendo, pero no limitados a, por ejemplo, lípidos, carbohidratos u otras moléculas orgánicas que son producidas por una o más moléculas genéticamente codificables (por ejemplo, una o más enzimas o vías enzimáticas) o similares.

Los términos "polinucleótido" y "ácido nucleico" se refieren a desoxirribonucleótidos o ribonucleótidos y polímeros (por ejemplo, oligonucleótidos, polinucleótidos, etc.) de los mismos en forma de cadena sencilla o doble. Estos términos incluyen, pero no se limitan a, ADN de cadena sencilla, doble o triple, ADN genómico, ADNc, ARN, híbridos de ADN-ARN, polímeros que comprenden bases de purina y pirimidina y/u otras bases nucleotídicas naturales, modificadas química o bioquímicamente, no naturales o derivatizadas. Los siguientes son ejemplos no limitantes de polinucleótidos: genes, fragmentos de genes, fragmentos cromosómicos, EST, exones, intrones, ARNm, ARNt, ARNr, ribozimas, ADNc, polinucleótidos recombinantes, polinucleótidos ramificados, plásmidos, vectores, ADN aislado de cualquier secuencia, ARN aislado de cualquier secuencia, sondas de ácidos nucleicos y cebadores. En algunas realizaciones, los polinucleótidos comprenden nucleótidos modificados, como nucleótidos metilados y análogos de nucleótidos, uracilo, otros azúcares y grupos de enlace como fluororibosa y tioato, y/o ramificaciones de nucleótidos. En algunas realizaciones alternativas, la secuencia de nucleótidos está interrumpida por componentes no nucleotídicos.

A menos que se limite específicamente, el término abarca ácidos nucleicos que contienen análogos conocidos de nucleótidos naturales que tienen propiedades de unión similares a las del ácido nucleico de referencia y se metabolizan de manera similar a los nucleótidos de origen natural. A menos que se indique lo contrario, una secuencia particular de ácidos nucleicos también abarca implícitamente variantes conservadoramente modificadas de la misma (por ejemplo, sustituciones degeneradas de codones) y secuencias complementarias, así como la secuencia explícitamente indicada. Específicamente, las sustituciones degeneradas de codones pueden lograrse generando secuencias en las que la tercera posición de uno o más codones seleccionados (o todos) se sustituye con residuos de base mixta y/o desoxiinosina (Batzer et al. (1991) Nucleic Acid Res. 19:5081; Ohtsuka et al. (1985) J. Biol. Chem. 260:2605-2608; Rossolini et al. (1994) Mol. Cell. Probes 8:91-98). El término ácido nucleico se usa indistintamente con, por ejemplo, oligonucleótido, polinucleótido, ADNc y ARNm.

Los términos "proteína", "polipéptido" y "péptido" se usan indistintamente para designar un polímero de por lo menos dos aminoácidos enlazados covalentemente por un enlace amida, independientemente de su longitud o modificación postraduccional (por ejemplo, glicosilación, fosforilación, lipidación, miristilación, ubiquitinación, etc.). En algunos casos, el polímero tiene por lo menos aproximadamente 30 residuos de aminoácidos, y habitualmente por lo menos aproximadamente 50 residuos de aminoácidos. Más típicamente, contienen por lo menos aproximadamente 100 residuos de aminoácidos. Los términos incluyen composiciones convencionalmente consideradas como fragmentos de proteínas o péptidos de longitud completa. En esta definición se incluyen los aminoácidos D y L, y las mezclas de aminoácidos D y L. Los polipéptidos descritos en la presente no se limitan a los aminoácidos codificados genéticamente. De hecho, además de los aminoácidos codificados genéticamente, los polipéptidos descritos en la presente pueden estar compuestos, total o parcialmente, de aminoácidos de origen natural y/o sintéticos no codificados. En algunas realizaciones, un polipéptido es una porción del polipéptido ancestral o parental de longitud completa, que contiene adiciones o deleciones (por ejemplo, huecos) de aminoácidos o sustituciones en comparación con la secuencia de aminoácidos del polipéptido parental de longitud completa, a la vez que conserva la actividad funcional (por ejemplo, la actividad catalítica).

Como se usa en la presente, el término "celulasa" se refiere a una categoría de enzimas capaces de hidrolizar la celulosa (β-1,4-glucano o enlaces β-D-glucosídicos) en cadenas de celulosa más cortas, oligosacáridos, celobiosa y/o glucosa. En algunas realizaciones, el término "celulasa" abarca beta-glucosidasas, endoglucanasas, celobiohidrolasas, celobiosa deshidrogenasas, endoxilanasas, beta-xilosidasas, arabinofuranosidasas, alfa-glucuronidasas, acetilxilano esterasas, feruloil esterasas, y/o alfa-glucuronil esterasas. En algunas realizaciones, el término "celulasa" engloba enzimas que hidrolizan hemicelulosa, incluyendo pero no limitadas a endoxilanasas, beta-xilosidasas, arabinofuranosidasas, alfa-glucuronidasas, acetilxilano esterasa, feruloil esterasa y alfa-glucuronil esterasa. Una "célula fúngica productora de celulasa" es una célula fúngica que expresa y secreta por lo menos una enzima hidrolizante de celulosa. En algunas realizaciones, las células fúngicas productoras de celulasa expresan y secretan una mezcla de enzimas hidrolizadoras de celulosa. "Celulolítico", "hidrolizante de celulosa", "degradante de celulosa" y términos similares se refieren a enzimas como endoglucanasas y celobiohidrolasas (estas últimas también denominadas "exoglucanasas") que actúan sinérgicamente para descomponer la celulosa en di- u oligosacáridos solubles como la celobiosa, que luego son hidrolizados a glucosa por la beta-glucosidasa. En algunas realizaciones, la celulasa es una celulasa recombinante seleccionada entre β-glucosidasas (BGLs), celobiohidrolasas de tipo 1 (CBH1s), celobiohidrolasas de tipo 2 (CBH2s), glucósido hidrolasa 61s (GH61s), y/o endoglucanasas (EGs). En algunas realizaciones, la celulasa es una celulasa recombinante *de Myceliophthora* seleccionada entre β-glucosidasas (BGLs), celobiohidrolasas de Tipo 1 (CBH1s), celobiohidrolasas de Tipo 2 (CBH2s), glucósido hidrolasa 61s (GH61s), y/o endoglucanasas (EGs). En algunas realizaciones adicionales, la celulasa es una celulasa recombinante seleccionada entre EG1b, EG2, EG3, EG4, EG5, EG6, CBH1a, CBH1b, CBH2a, CBH2b, GH61a, y/o BGL.

El término "secuencia" se usa en la presente para referirse al orden y la identidad de cualquier secuencia

biológica incluyendo, pero no limitado a, un genoma completo, un cromosoma completo, un segmento cromosómico, una colección de secuencias genéticas de genes que interactúan, un gen, una secuencia de ácido nucleico, una proteína, un polisacárido, etc. En algunos contextos, una secuencia se refiere al orden e identidad de los residuos de aminoácidos en una proteína (es decir, una secuencia de proteína o cadena de caracteres de proteína) o al orden e identidad de los nucleótidos en un ácido nucleico (es decir, una secuencia de ácido nucleico o cadena de caracteres de ácido nucleico). Una secuencia puede representarse mediante una cadena de caracteres. Una "secuencia de ácido nucleico" se refiere al orden e identidad de los nucleótidos que componen un ácido nucleico. Una "secuencia de proteína" se refiere al orden e identidad de los aminoácidos que componen una proteína o péptido.

"Codón" se refiere a una secuencia específica de tres nucleótidos consecutivos que forma parte del código genético y que especifica un aminoácido particular en una proteína o inicia o detiene la síntesis de proteínas.

"Secuencia nativa" o "secuencia de tipo salvaje" se refiere a un polinucleótido o polipéptido aislado de una fuente natural. Dentro de la "secuencia nativa" se incluyen las formas recombinantes de un polipéptido o polinucleótido nativo que tienen una secuencia idéntica a la forma nativa.

El término "gen" se usa en sentido amplio para referirse a cualquier segmento de ADN u otro ácido nucleico asociado a una función biológica. Por tanto, los genes incluyen secuencias codificantes y, opcionalmente, las secuencias reguladoras necesarias para su expresión. Los genes también incluyen opcionalmente segmentos de ácido nucleico no expresados que, por ejemplo, forman secuencias de reconocimiento para otras proteínas. Los genes pueden obtenerse de una variedad de fuentes, incluyendo la clonación a partir de una fuente de interés o la síntesis a partir de información de secuencias conocidas o predichas, y pueden incluir secuencias diseñadas para tener los parámetros deseados.

Un "motivo" se refiere a un patrón de subunidades en o entre moléculas biológicas. Por ejemplo, el término "motivo" puede usarse en referencia a un patrón de subunidades de la molécula biológica no codificada o a un patrón de subunidades de una representación codificada de una molécula biológica.

El término "cromosoma" se usa en referencia a una estructura organizada de ADN y proteínas asociadas que se encuentra en las células y que comprende una única pieza de ADN enrollado que incluye muchos genes, elementos reguladores y otras secuencias de nucleótidos. El término también se usa en referencia a la secuencia de ADN de la estructura.

"Cribado" se refiere al proceso en el que se determinan una o más propiedades de una o más biomoléculas. Por ejemplo, los procesos típicos de cribado incluyen aquellos en los que se determinan una o más propiedades de uno o más miembros de una o más bibliotecas. Un "sistema de expresión" es un sistema para expresar una proteína o péptido codificado por un gen u otro ácido nucleico.

"Célula huésped" o "célula huésped recombinante" se refiere a una célula que comprende por lo menos una molécula de ácido nucleico recombinante. Así, por ejemplo, en algunas realizaciones, las células huésped recombinantes expresan genes que no se encuentran en la forma nativa (es decir, no recombinante) de la célula.

"Evolución dirigida", "evolución guiada" o "evolución artificial" se refiere a procesos in vitro o in vivo de cambio artificial de una o más secuencias de biomoléculas (o una cadena de caracteres que representa esa secuencia) mediante selección artificial, recombinación u otra manipulación. En algunas realizaciones, la evolución dirigida se produce en una población reproductiva en la que hay (1) variedades de individuos, siendo algunas variedades (2) heredables, de las cuales algunas variedades (3) difieren en idoneidad. El éxito reproductivo se determina por el resultado de la selección de una propiedad predeterminada, como una propiedad beneficiosa. La población reproductora puede ser, por ejemplo, una población física o una población virtual en un sistema informático.

En ciertas realizaciones, los métodos de evolución dirigida generan bibliotecas de variantes de proteínas recombinando genes que codifican variantes de una biblioteca de variantes de proteínas parentales. Los métodos pueden emplear oligonucleótidos que contengan secuencias o subsecuencias para codificar las proteínas de una biblioteca de variantes parentales. Algunos de los oligonucleótidos de la biblioteca de variantes parentales pueden estar estrechamente relacionados, diferenciándose sólo en la elección de codones para aminoácidos alternativos seleccionados para ser variados por recombinación con otras variantes. El método puede realizarse durante uno o varios ciclos hasta que se obtengan los resultados deseados. Si se usan múltiples ciclos, cada uno de ellos implica un paso de cribado para identificar qué variantes con un rendimiento aceptable se usarán en un ciclo de recombinación posterior.

Los términos "trasposición" y "trasposición de genes" se refieren a métodos de evolución dirigida para introducir diversidad mediante la recombinación de una colección de fragmentos de polinucleótidos parentales a través de una serie de ciclos de extensión de cadena. En ciertas realizaciones, uno o más de los ciclos de extensión de cadena es autocebante, es decir, se realiza sin la adición de cebadores distintos de los propios fragmentos. Cada ciclo implica el apareamiento de fragmentos de cadena sencilla mediante hibridación, la posterior elongación de los

fragmentos apareados mediante la extensión de la cadena y la desnaturalización. En el transcurso de la trasposición, se expone típicamente una cadena de ácido nucleico creciente a múltiples compañeros de apareamiento diferentes en un proceso denominado a veces "cambio de plantilla". Como se usa en la presente, "cambio de plantilla" se refiere a la capacidad de cambiar un dominio de ácido nucleico de un ácido nucleico con un segundo dominio de un segundo ácido nucleico (es decir, el primer y el segundo ácido nucleico sirven como plantillas en el procedimiento de trasposición).

El cambio de plantilla produce con frecuencia secuencias quiméricas, que resultan de la introducción de cruces entre fragmentos de orígenes diferentes. Los cruces se crean a través de recombinaciones de cambio de plantilla durante los múltiples ciclos de apareamiento, extensión y desnaturalización. Por tanto, la transposición lleva típicamente a la producción de secuencias de polinucleótidos variantes. En algunas realizaciones, las secuencias variantes comprenden una "biblioteca" de variantes. En algunas realizaciones de estas bibliotecas, las variantes contienen segmentos de secuencia de dos o más polinucleótidos parentales.

Cuando se emplean dos o más polinucleótidos parentales, los polinucleótidos parentales individuales son suficientemente homólogos para que los fragmentos de diferentes parentales hibriden en las condiciones de apareamiento empleadas en los ciclos de trasposición. En algunas realizaciones, la trasposición permite la recombinación de polinucleótidos parentales que tienen una homología relativamente limitada. A menudo, los polinucleótidos parentales individuales tienen dominios distintos y/o únicos y/u otras características de secuencia de interés. Cuando se usan polinucleótidos parentales con características de secuencia distintas, la trasposición puede producir polinucleótidos variantes muy diversos.

En la técnica se conocen varias técnicas de trasposición. Consultar, por ejemplo, las Patentes de Estados Unidos Nº 6,917,882, 7,776,598, 8,029,988, 7,024,312, y 7,795,030.

Un "fragmento" es cualquier porción de una secuencia de nucleótidos o aminoácidos. Los fragmentos pueden producirse usando cualquier método adecuado conocido en la técnica, incluyendo pero no limitados a, la escisión de una secuencia de polipéptidos o polinucleótidos. En algunas realizaciones, los fragmentos se producen usando nucleasas que escinden polinucleótidos. En algunas realizaciones adicionales, los fragmentos se generan usando técnicas de síntesis química y/o biológica. En algunas realizaciones, los fragmentos comprenden subsecuencias de por lo menos una secuencia parental, generadas mediante elongación parcial de cadena de ácido o ácidos nucleicos complementarios.

"Polipéptido parental", "polinucleótido parental", "ácido nucleico parental" y "parental" se usan generalmente para referirse al polipéptido de tipo salvaje, polinucleótido de tipo salvaje o una variante usada como punto de partida en un procedimiento de generación de diversidad, como una evolución dirigida. En algunas realizaciones, el propio progenitor se produce mediante trasposición u otro procedimiento de generación de diversidad. En algunas realizaciones, los mutantes usados en la evolución dirigida están directamente relacionados con un polipéptido parental. En algunas realizaciones, el polipéptido parental es estable cuando se expone a condiciones extremas de temperatura, pH y/o solventes y puede servir como base para generar variantes para la trasposición. En algunas realizaciones, el polipéptido parental no es estable en condiciones extremas de temperatura, pH y/o solvente, y el polipéptido parental evoluciona para generar variantes robustas.

Un "ácido nucleico parental" codifica un polipéptido parental.

"Mutante", "variante" y "secuencia variante", como se usan en la presente, se refieren a una secuencia biológica que difiere en algún aspecto de una secuencia estándar o de referencia. La diferencia puede denominarse "mutación". En algunas realizaciones, un mutante es un aminoácido (es decir, polipéptido) o secuencia de polinucleótidos que ha sido alterada por al menos una sustitución, inserción, cruce, deleción y/u otra operación genética. A efectos de la presente divulgación, los mutantes y variantes no están limitados a un método particular por el que se generan. En algunas realizaciones, una secuencia mutante o variante tiene actividades o propiedades aumentadas, disminuidas o sustancialmente similares, en comparación con la secuencia parental. En algunas realizaciones, el polipéptido variante comprende uno o más residuos de aminoácidos que han sido mutados, en comparación con la secuencia de aminoácidos del polipéptido de tipo salvaje (por ejemplo, un polipéptido parental). En algunas realizaciones, uno o más residuos de aminoácidos del polipéptido se mantienen constantes, son invariantes, o no están mutados en comparación con un polipéptido parental en los polipéptidos variantes que componen la pluralidad. En algunas realizaciones, el polipéptido parental se usa como base para generar variantes con estabilidad, actividad u otra propiedad mejoradas.

"Mutagénesis" es el proceso de introducir una mutación en una secuencia estándar o de referencia, como un ácido nucleico parental o un polipéptido parental.

Una "biblioteca" o "población" se refiere a una colección de por lo menos dos moléculas, cadenas de caracteres y/o modelos diferentes, como secuencias de ácidos nucleicos (por ejemplo, genes, oligonucleótidos, etc.) o productos de expresión (por ejemplo, enzimas u otras proteínas) de los mismos. Una biblioteca o población

generalmente incluye varias moléculas diferentes. Por ejemplo, una biblioteca o población incluye típicamente por lo menos aproximadamente 10 moléculas diferentes. Las bibliotecas grandes incluyen típicamente por lo menos aproximadamente 100 moléculas diferentes, más típicamente por lo menos aproximadamente 1000 moléculas diferentes. En algunas aplicaciones, la biblioteca incluye por lo menos aproximadamente 10000 moléculas diferentes o más. En ciertas realizaciones, la biblioteca contiene una variante numérica o ácidos nucleicos quiméricos o proteínas producidos por un procedimiento de evolución dirigida.

Dos ácidos nucleicos se "recombinan" cuando las secuencias de cada uno de los dos ácidos nucleicos se combinan en un ácido nucleico de progenie. Dos secuencias se recombinan "directamente" cuando ambos ácidos nucleicos son sustratos para la recombinación.

"Selección" se refiere al proceso en el que una o más biomoléculas se identifican como poseedoras de una o más propiedades de interés. Así, por ejemplo, puede cribarse una biblioteca para determinar una o más propiedades de uno o más miembros de la biblioteca. Si se identifica que uno o más de los miembros de la biblioteca poseen una propiedad de interés, se seleccionan. La selección puede incluir el aislamiento de un miembro de la biblioteca, pero esto no es necesario. Además, la selección y el cribado pueden ser, y a menudo son, simultáneos.

Una "variable dependiente" representa un resultado o efecto, o se prueba para ver si es el efecto. Las "variables independientes" representan las entradas o causas, o se prueban para ver si son la causa. Puede estudiarse una variable dependiente para ver si varía, y en qué medida, a media que varían las variables independientes.

En el modelo lineal estocástico simple

$$y_i = a + bx_i + e_i$$

em donde el término y_i es el valor i-ésimo de la variable dependiente y x_i es el valor i-ésimo de la variable independiente. El término e_i se conoce como el "error" y contiene la variabilidad de la variable dependiente no explicada por la variable independiente.

Una variable independiente también se conoce como "variable predictora", "regresor", "variable controlada", "variable manipulada", "variable explicativa" o "variable de entrada".

"Ortogonal/ortogonalidad" se refiere a una variable independiente que no está correlacionada con otras variables independientes en un modelo u otra relación.

El término "modelo secuencia-actividad" se refiere a cualquier modelo matemático que describa la relación entre actividades, características o propiedades de moléculas biológicas, por un lado, y varias secuencias biológicas, por otro.

El término "cadena de caracteres codificada" se refiere a una representación de una molécula biológica que conserva información secuencial/estructural relativa a dicha molécula. En algunas realizaciones, la cadena de caracteres codificada contiene información sobre mutaciones de secuencia en una biblioteca de variantes. Las cadenas de caracteres codificadas de biomoléculas junto con la información de actividad de las biomoléculas pueden usarse como conjunto de entrenamiento para un modelo de actividad de secuencia. Las propiedades no secuenciales de las biomoléculas pueden almacenarse o asociarse de otro modo con cadenas de caracteres codificadas para las biomoléculas.

"Secuencia de referencia" es una secuencia a partir de la cual se efectúa la variación de la secuencia. En algunos casos, una "secuencia de referencia" se usa para definir las variaciones. Dicha secuencia puede ser una que un modelo ha predicho que tendrá el valor más alto (o uno de los valores más altos) de la actividad deseada. En otro caso, la secuencia de referencia puede ser la de un miembro de una biblioteca original de variantes de proteínas. En ciertas realizaciones, una secuencia de referencia es la secuencia de una proteína o ácido nucleico parental.

"Conjunto de entrenamiento" se refiere a un conjunto de datos u observaciones de secuencia-actividad a los que se ajustan uno o más modelos y sobre los que se construyen. Por ejemplo, para un modelo de secuencia-actividad de proteínas, un conjunto de entrenamiento comprende secuencias de residuos para una biblioteca de variantes de proteínas inicial o mejorada. Típicamente, estos datos incluyen información completa o parcial de la secuencia de residuos, junto con un valor de actividad para cada proteína de la biblioteca. En algunos casos, múltiples tipos de actividades (por ejemplo, datos de constante de velocidad y datos de estabilidad térmica) se proporcionan juntos en el conjunto de entrenamiento. A veces, la actividad es una propiedad beneficiosa.

El término "observación" se refiere a la información sobre una proteína u otra entidad biológica que puede usarse en un conjunto de entrenamiento para generar un modelo, como un modelo de actividad de secuencias. El término "observación" puede referirse a cualquier molécula biológica secuenciada y ensayada, incluyendo las variantes de proteínas. En ciertas realizaciones, cada observación es un valor de actividad y una secuencia asociada

para una variante en una biblioteca. En general, cuantas más observaciones se empleen para crear un modelo de secuencia-actividad, mejor será la potencia predictiva de dicho modelo.

Como se usa en la presente, se pretende que el término "propiedad beneficiosa" se refiera a una característica fenotípica u otra característica identificable que confiere algún beneficio a una proteína o a una composición de materia o proceso asociado con la proteína. Los ejemplos de propiedades beneficiosas incluyen un aumento o disminución, en comparación con una proteína parental, de las propiedades catalíticas, las propiedades de unión, la estabilidad cuando se expone a temperaturas extremas, pH, etc., la sensibilidad a estímulos, la inhibición y similares de una proteína variante. Otras propiedades beneficiosas pueden incluir un perfil alterado en respuesta a un estímulo particular. A continuación se exponen otros ejemplos de propiedades beneficiosas. Los valores de las propiedades beneficiosas pueden usarse como valores de actividad en las observaciones usadas en un conjunto de entrenamiento para un modelo de actividad de secuencia.

La "secuenciación de próxima generación" o "secuenciación de alto rendimiento" son técnicas de secuenciación que paralelizan el proceso de secuenciación, produciendo miles o millones de secuencias a la vez. Algunos ejemplos de métodos adecuados de secuenciación de próxima generación incluyen, entre otros, la secuenciación en tiempo real de una sola molécula (por ejemplo, Pacific Biosciences, Menlo Park, California), la secuenciación por semiconductores iónicos (por ejemplo, Ion Torrent, South San Francisco, California), la pirosecuenciación (por ejemplo, 454, Branford, Connecticut), secuenciación por ligadura (por ejemplo, secuenciación SOLid de Life Technologies, Carlsbad, California), secuenciación por síntesis y terminador reversible (por ejemplo, Illumina, San Diego, California), tecnologías de imagenología de ácidos nucleicos como la microscopía electrónica de transmisión, y similares. En la descripción detallada de la presente divulgación se describen descripciones adicionales de técnicas ejemplares.

"Potencia predictiva" se refiere a la capacidad de un modelo para predecir correctamente los valores de una variable dependiente para datos bajo varias condiciones. Por ejemplo, la potencia predictiva de un modelo de actividad secuencial se refiere a la capacidad del modelo para predecir la actividad a partir de información secuencial.

La "validación cruzada" se refiere a un método para comprobar la generalizabilidad de la capacidad de un modelo para predecir un valor de interés (es decir, el valor de la variable dependiente). El método prepara un modelo usando un conjunto de datos y comprueba el error del modelo usando un conjunto de datos diferente. El primer conjunto de datos se considera un conjunto de entrenamiento, y el segundo, un conjunto de validación.

"Varianza sistemática" se refiere a diferentes descriptores de un artículo o conjunto de artículos que se cambian en diferentes combinaciones.

"Datos sistemáticamente variados" se refiere a los datos producidos, derivados o resultantes de diferentes descriptores de un artículo o conjunto de artículos que se cambian en diferentes combinaciones. Muchos descriptores diferentes pueden cambiarse al mismo tiempo, pero en diferentes combinaciones. Por ejemplo, los datos de actividad recopilados de polipéptidos en los que se han cambiado combinaciones de aminoácidos son datos sistemáticamente variados.

El término "secuencias sistemáticamente variadas" se refiere a un conjunto de secuencias en las que cada residuo se observa en múltiples contextos. En principio, el nivel de variación sistemática puede cuantificarse por el grado en que las secuencias son ortogonales entre sí (es decir, máximamente diferentes en comparación con la media).

El término "alternante" se refiere a la introducción de múltiples tipos de residuos de aminoácidos en una posición específica en las secuencias de las variantes de proteínas de la biblioteca optimizada.

Los términos "regresión" y "análisis de regresión" se refieren a técnicas usadas para comprender qué variables independientes están relacionadas con la variable dependiente y explorar las formas de estas relaciones. En circunstancias restringidas, el análisis de regresión puede usarse para inferir relaciones causales entre las variables independientes y dependientes. Es una técnica estadística para estimar las relaciones entre variables. Incluye muchas técnicas de modelización y análisis de varias variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes. Más específicamente, el análisis de regresión ayuda a comprender cómo cambia el valor típico de la variable dependiente cuando se varía cualquiera de las variables independientes, mientras que las demás variables independientes se mantienen fijas. Las técnicas de regresión pueden usarse para generar modelos de actividad de secuencia a partir de conjuntos de entrenamiento que comprenden múltiples observaciones, que pueden contener información sobre la secuencia y la actividad.

Mínimos cuadrados parciales o PLS es una familia de métodos que encuentra un modelo de regresión lineal proyectando las variables predichas (por ejemplo, actividades) y las variables observables (por ejemplo, secuencias) a un nuevo espacio. PLS también se conoce como proyección a estructuras latentes. Tanto los datos X (variables independientes) como Y (variables dependientes) se proyectan a nuevos espacios. El PLS se usa para encontrar las

relaciones fundamentales entre dos matrices (X e Y). Se usa un enfoque de variable latente para modelar las estructuras de covarianza en los espacios X e Y. Un modelo PLS intentará encontrar la dirección multidimensional en el espacio X que explique la máxima dirección de varianza multidimensional en el espacio Y. La regresión PLS es particularmente adecuada cuando la matriz de predictores tiene más variables que observaciones y cuando existe multicolinealidad entre los valores X.

Un "descriptor" se refiere a algo que sirve para describir o identificar un artículos. Por ejemplo, los caracteres de una cadena de caracteres pueden ser descriptores de los aminoácidos de un polipéptido que está representado por la cadena de caracteres.

En un modelo de regresión, la variable dependiente se relaciona con las variables independientes mediante una suma de términos. Cada término incluye un producto de una variable independiente y un coeficiente de regresión asociado. En el caso de un modelo de regresión puramente lineal, los coeficientes de regresión vienen dados por **β** en la siguiente forma de expresión:

$$y = \beta_{i1}x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x_i}^T\boldsymbol{\beta} + \varepsilon_i$$

donde y_i es la variable dependiente, las x_i son las variables independientes, ε_i es la variable de error, y T denota la transposición, es decir, el producto interior de los vectores $\mathbf{x_i}$ y **β**.

La "regresión de componentes principales" (RCP) se refiere a un análisis de regresión que usa el análisis de componentes principales para estimar los coeficientes de regresión. En la PCR, en lugar de aplicar directamente la regresión de la variable dependiente sobre las variables independientes, se usan los componentes principales de las variables independientes. Típicamente, la PCR sólo usa un subconjunto de los componentes principales en la regresión.

"Análisis de componentes principales" (PCA) se refiere a un procedimiento matemático que usa una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas denominadas componentes principales. El número de componentes principales es menor o igual que el número de variables originales. Esta transformación se define de tal manera que el primer componente principal tenga la mayor varianza posible (es decir, que represente la mayor parte posible de la variabilidad de los datos), y cada componente sucesivo tenga a su vez la mayor varianza posible con la restricción de que sea ortogonal a los componentes precedentes (es decir, que no esté correlacionado con ellos).

"Red neuronal" es un modelo que contiene un grupo interconectado de elementos de procesamiento o "neuronas" que procesan información usando un enfoque conexionista de la computación. Las redes neuronales se usan para modelar relaciones complejas entre entradas y salidas o para encontrar patrones en los datos. La mayoría de las redes neuronales procesan los datos de manera no lineal, distribuida y paralela. En la mayoría de los casos, una red neuronal es un sistema adaptativo que cambia su estructura durante una fase de aprendizaje. Las funciones se realizan colectivamente y en paralelo por los elementos de procesamiento, en lugar de existir una clara delimitación de las subtareas a las que se asignan las distintas unidades.

En general, una red neuronal consiste en una red de elementos de procesamiento simples que muestran un comportamiento global complejo determinado por las conexiones entre los elementos de procesamiento y los parámetros de los elementos. Las redes neuronales se usan con algoritmos diseñados para alterar la fuerza de las conexiones en la red para producir un flujo de señal deseado. La fuerza se modifica durante el entrenamiento o el aprendizaje.

"Bosque aleatorio" se refiere a una combinación de predictores de árboles de clasificación tal que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles del bosque. Un bosque aleatorio es un conjunto de aprendizaje que consiste en un conjunto de árboles de decisión no podados con una selección aleatoria de características en cada división del árbol de decisión. Un bosque aleatorio genera un gran número de árboles de clasificación, cada uno de los cuales vota por la clase más popular. A continuación, el bosque aleatorio clasifica una variable tomando la clase más votada de entre todos los predictores de árboles del bosque.

"Distribución de probabilidad a priori", o "a priori", de una cantidad incierta p es la distribución de probabilidad que expresa la incertidumbre sobre p antes de que se tengan en cuenta los datos de interés (por ejemplo, un conjunto de entrenamiento de secuencias de proteínas). La cantidad desconocida puede ser un parámetro, coeficiente, variable, variable latente o similar (por ejemplo, un coeficiente en un modelo de regresión múltiple).

"Distribución de probabilidad a posteriori", o "posterior", de una cantidad incierta p es la distribución de probabilidad que expresa la incertidumbre sobre p después de tener en cuenta los datos de interés.

El término "regresión lineal bayesiana" se refiere a un enfoque de la regresión lineal en el que el análisis estadístico se lleva a cabo en el contexto de la inferencia bayesiana. La creencia previa sobre el modelo de regresión lineal, incluyendo la función de distribución de probabilidad previa del parámetro del modelo, se combina con la función de verosimilitud de los datos según el teorema de Bayes para obtener la distribución de probabilidad posterior sobre los parámetros.

"Sobreajuste" se refiere a la condición que se produce cuando un modelo estadístico describe un error aleatorio o ruido en lugar de la relación subyacente. El sobreajuste se produce generalmente cuando un modelo es excesivamente complejo, por ejemplo, cuando tiene demasiados parámetros con respecto al número de observaciones. Un modelo que ha sido sobreajustado generalmente tendrá un rendimiento predictivo pobre, ya que puede exagerar fluctuaciones menores en los datos. En algunas realizaciones, se usa un modelo matemático para describir la relación entre una o más variables independientes (IV) y una variable dependiente (DV). El modelo puede escribirse como DV = Expresión algebraica de (IVs). Una "expresión algebraica" puede incluir variables, coeficientes, constantes y símbolos operativos, como los signos más y menos. $4x^2 + 3xy + 7y + 5$ es una expresión algebraica bivariante.

En algunas realizaciones, los "términos" de una expresión algebraica o de un modelo matemático son los elementos separados por los signos más o menos. En este contexto, el ejemplo anterior tiene cuatro términos, $4x^2$, $3xy$, $7y$ y 5. Los términos pueden consistir en variables y coeficientes ($4x^2$, $3xy$, y $7y$), o constantes (5). En las expresiones algebraicas, las variables pueden tomar varios valores para representar condiciones cambiantes de un sistema. Por ejemplo, puede ser una variable continua que represente la velocidad de un coche en marcha o una variable discreta con múltiples valores no continuos que representen tipos de aminoácidos. Una variable puede ser una variable de valor de bit que representa la presencia o ausencia de una entidad, por ejemplo, la presencia o ausencia de un residuo de un tipo específico en una posición específica. En la expresión algebraica anterior, las variables son x e y.

En algunos casos, los "términos" de una expresión pueden ser elementos de la expresión que están delimitados por otros signos, como la multiplicación.

"Coeficiente" se refiere a un valor escalar multiplicado por una variable dependiente o una expresión que contiene una variable dependiente. En el ejemplo anterior, los "coeficientes" son la parte numérica de los términos de una expresión algebraica. En $4x^2 + 3xy + 7y + 5$, el coeficiente del primer término es 4. El coeficiente del segundo término es 3, y el coeficiente del tercer término es 7. Si un término consiste sólo de variables, su coeficiente es 1.

Las "constantes" son los términos de la expresión algebraica que sólo contienen números. Es decir, son los términos sin variables. En la expresión $4x^2 + 3xy + 7y + 5$, el término constante es "5".

Un "término lineal" es un término con un grado de 1, o una única variable elevada a la potencia de 1. En el ejemplo anterior, el término $7y$ es un término lineal porque su grado es 1 (y^1 o simplemente y). Por el contrario, el término $4x^2$ es un término cuadrático porque la x tiene un grado de 2, y $3xy$ es un término cuadrático bivariante porque x e y tienen cada uno un grado de 1, y el producto conduce a un grado de 2.

En algunos lugares del texto, "término lineal" y "término de no interacción" se usan indistintamente en la presente para referirse a un término de un modelo de regresión que comprende el producto de una única variable independiente y un coeficiente asociado, en donde la única IV representa la presencia/ausencia de un único residuo.

En algunas realizaciones, "término no lineal", "término de producto cruzado" y "término de interacción" se usan indistintamente en esta divulgación cuando se refieren a un término de un modelo de regresión que comprende el producto de dos o más variables independientes y un coeficiente asociado. En términos más generales, los "términos no lineales" se usan para indicar términos con un grado mayor o menor que 1, por ejemplo, una función de potencia o una función exponencial de la variable independiente. Algunos ejemplos de términos no lineales incluyen xy, x^2, $x^{1/3}$, x^y, y e^x. Por tanto, en algunos lugares del texto, "término no lineal" se refiere a un sentido más amplio que un término que incluye el producto de dos variables independientes.

En algunas realizaciones, un término de interacción puede implementarse como un término que incluye una función no lineal de dos o más IV, por ejemplo, la función producto, función potencia o función exponencial de dos o más IV, cada IV representando la presencia de un residuo de un tipo específico en una posición específica. Por ejemplo, en $y = ax_1 + bx_2 + cx_1x_2$, las variables x_1 y x_2 pueden representar la presencia/ausencia de dos residuos concretos en una posición particular, y el término cx_1x_2 es un término de interacción que representa el efecto de la interacción de los dos residuos particulares. En otras realizaciones, un término de interacción puede implementarse como un término que incluye una única IV que representa la interacción de dos o más residuos. Por ejemplo, en $y = ax_1 + bx_2 + cz$, las variables x_1 y x_2 pueden representar la presencia/ausencia de dos residuos particulares en una posición particular, y el término cz es un término de interacción que representa el efecto de la interacción de los dos residuos particulares. En este último ejemplo, el término de interacción cz no es un término de producto cruzado. Aunque técnicamente cz es un término lineal, no se marca así en la presente para evitar confusiones con los términos

lineales y no interactivos ax_1 y bx_2. Como se usa en la divulgación, el término "modelo lineal" se refiere a modelos que incluyen sólo términos lineales. Por el contrario, el término "modelo no lineal" se refiere a modelos que incluyen tanto términos lineales como no lineales. En algunas realizaciones, los modelos no lineales incluyen términos de interacción implementados como términos de productos cruzados.

De manera más general, un modelo lineal o un sistema lineal satisface el principio de superposición y la homogeneidad de grado 1. El principio de superposición establece que, para todos los sistemas lineales, la respuesta neta en un lugar y un momento dados provocada por dos o más estímulos es la suma de las respuestas que habría provocado cada estímulo por separado. Esto también se conoce como aditividad. Si la entrada A produce la respuesta X y la entrada B produce la respuesta Y, entonces la entrada (A+B) produce la respuesta (X+Y). La homogeneidad de grado 1 se refiere a cualquier modelo cuya salida o variable dependiente (VD) cambia proporcionalmente a su entrada o variable independiente. Por el contrario, un "modelo no lineal" es un modelo que no satisface el principio de superposición ni la homogeneidad de grado 1.

Por "subunidades que interactúan" se entiende dos o más subunidades de una secuencia que tienen un efecto sinérgico sobre la actividad modelada de la secuencia, siendo el efecto sinérgico siendo independiente y diferente de los efectos individuales de las subunidades sobre la actividad modelada.

El término "modelo base" se usa en referencia a un modelo de secuencia-actividad proporcionado al principio de un proceso de mejora de un modelo.

El término "modelo actualizado" se usa en referencia a un modelo de secuencia-actividad que se deriva directa o indirectamente de un modelo base, que tiene una potencia predictiva mejorada en comparación con el modelo base y/u otro modelo del que se deriva.

Una "función de verosimilitud" o " verosimilitud" de un modelo es una función de los parámetros de un modelo estadístico. La verosimilitud de un conjunto de valores de parámetros dados unos resultados observados es igual a la probabilidad de esos resultados observados dados esos valores de parámetros, es decir, $L(\theta|x) = P(x|\theta)$.

Las "simulaciones de Montecarlo" son simulaciones que se basan en un gran número de muestreos aleatorios para obtener resultados numéricos que simulen un fenómeno real. Por ejemplo, extraer un gran número de variables uniformes pseudoaleatorias del intervalo (0,1], y asignar valores menores o iguales a 0,50 como cara y mayores que 0,50 como cruz, es una simulación de Montecarlo del comportamiento de lanzar repetidamente una moneda al aire.

Un "algoritmo Metrópolis" o "algoritmo Metrópolis-Hastings" es un método de Monte Carlo con cadena de Markov (MCMC) para obtener una secuencia de muestras aleatorias de una distribución de probabilidad para la que el muestreo directo es difícil. Esta secuencia de muestreo puede usarse para aproximar la distribución (es decir, para generar un histograma) o para calcular una integral (como un valor esperado). Los algoritmos Metropolis-Hastings y otros algoritmos MCMC se usan generalmente para el muestreo de distribuciones multidimensionales, especialmente cuando el número de dimensiones es elevado. El objetivo del algoritmo Metropolis-Hastings es generar asintóticamente estados x de acuerdo con una distribución deseada $P(x)$ y usa un proceso estocástico para cumplirlo. La idea del algoritmo es condicionar el proceso estocástico de tal manera que converja asintóticamente a la distribución única $P(x)$.

Una "cadena de Markov" es una secuencia de variables aleatorias X_1, X_2, X_3 ... con la propiedad de Markov. En otras palabras, dado el estado actual, los estados futuros y pasados son independientes. Formalmente,

$$\mathbf{Pr(X_{n+1} = x|X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = Pr(X_{n+1} = x|X_n = x_n)}$$

Los posibles valores de X_i forman un conjunto contable S denominado espacio de estados de la cadena. Un sistema de "cadena de Markov" es un sistema matemático que experimenta transiciones de un estado a otro, entre un número finito o contable de estados posibles. Se trata de un proceso aleatorio que habitualmente se caracteriza por carecer de memoria: el estado siguiente sólo depende del estado actual y no de la secuencia de acontecimientos que lo precedieron.

El "criterio de información de Akaike" (AIC) es una medida de la bondad relativa de ajuste de un modelo estadístico, y se usa a menudo como criterio de selección de modelos entre un conjunto finito de modelos. El AIC se basa en el concepto de entropía de la información y ofrece una medida relativa de la información que se pierde cuando se usa un modelo dado para describir la realidad. Puede decirse que describe el equilibrio entre el sesgo y la varianza en la construcción del modelo, o en términos generales, entre la precisión y la complejidad del modelo. El AIC puede calcularse como $AIC = -2\log_e L + 2k$, donde L es la máxima probabilidad de la función y k es el número de parámetros libres del modelo que hay que estimar.

El "criterio de Información bayesiano" es un criterio para la selección de modelos entre un conjunto finito de modelos, y está estrechamente relacionado con el AIC. El BIC puede calcularse como $BIC = -2\log_e L + k\log_e(n)$, en

donde n es el número de observaciones de datos. A medida que aumenta el número de observaciones, el BIC penaliza a menudo más el número extra de parámetros libres que el AIC.

Un "algoritmo genético" es un proceso que imita los procesos evolutivos. Los algoritmos genéticos (AG) se usan en una gran variedad de campos para resolver problemas que no están completamente caracterizados o que son demasiado complejos para permitir una caracterización completa, pero para los que se dispone de alguna evaluación analítica. Es decir, los AG se usan para resolver problemas que pueden evaluarse mediante alguna medida cuantificable del valor relativo de una solución (o por lo menos del valor relativo de una solución potencial en comparación con otra). En el contexto de la presente divulgación, un algoritmo genético es un proceso para seleccionar o manipular cadenas de caracteres en un ordenador, típicamente cuando la cadena de caracteres corresponde a una o más moléculas biológicas (por ejemplo, ácidos nucleicos, proteínas o similares).

El término "operación genética" (u "OG") se refiere a operaciones genéticas biológicas y/o computacionales, en donde todos los cambios en cualquier población de cualquier tipo de cadenas de caracteres (y, por tanto, en cualquier propiedad física de los objetos físicos codificados por tales cadenas) pueden describirse como resultado de la aplicación aleatoria y/o predeterminada de un conjunto finito de funciones algebraicas lógicas. Los ejemplos de GO incluyen, entre otros, la multiplicación, el cruce, la recombinación, la mutación, la ligadura, la fragmentación, etc.

"Modelo de conjunto" es un modelo cuyos términos incluyen todos los términos de un grupo de modelos, en el que los coeficientes de los términos del modelo de conjunto se basan en los coeficientes ponderados de los términos correspondientes de los modelos individuales del grupo. La ponderación de los coeficientes se basa en la potencia predictiva y/o la idoneidad de los modelos individuales.

II. GENERACIÓN DE BIBLIOTECAS MEJORADAS DE VARIANTES DE PROTEÍNAS

En un enfoque de evolución guiada para explorar secuencias de proteínas, se usan modelos de secuencia-actividad para guiar la generación de variantes de proteínas. Un aspecto de la divulgación proporciona varios métodos para preparar modelos de secuencia-actividad que se basan en bibliotecas de proteínas y pueden usarse para buscar bibliotecas de proteínas nuevas y mejoradas. Esta sección proporciona primero una visión general del proceso de búsqueda de proteínas nuevas y mejoradas, y luego proporciona detalles adicionales sobre cuestiones relacionadas con la selección de una biblioteca de partida, la construcción de un modelo de secuencia-actividad, y el uso del modelo para guiar la exploración de nuevas proteínas.

Esta divulgación proporciona ejemplos ilustrativos que implican secuencias de residuos de aminoácidos y actividades de proteínas, pero se entiende que el enfoque descrito en la presente también puede implementarse para otras secuencias y actividades biológicas. Por ejemplo, en varias realizaciones, una secuencia puede ser un genoma completo, un cromosoma completo, un segmento de cromosoma, una colección de secuencias génicas para genes que interactúan, un gen, una secuencia de ácido nucleico, una proteína, un polisacárido, etc. En una o más realizaciones, las subunidades de las secuencias pueden ser cromosomas, segmentos de cromosomas, haplotipos, genes, nucleótidos, codones, mutaciones, aminoácidos, carbohidratos mono, di, tri u oligoméricos, etc.

Típicamente, al comienzo de una ronda particular de evolución dirigida de secuencias, se obtiene un conjunto de entrenamiento de variantes de proteínas secuenciadas y ensayadas. Una ronda dada de evolución dirigida produce un número de proteínas variantes que varían en una o más mutaciones con respecto al péptido o péptidos parentales usados al principio de la ronda de evolución dirigida. Los péptidos variantes producidos durante una ronda de evolución dirigida se someten a ensayos de actividad. Aquellos péptidos que tengan la actividad deseada y/o una actividad mejorada en comparación con el péptido o péptidos parentales se seleccionan para su uso en por lo menos otra ronda de evolución dirigida.

Las variantes de proteínas secuenciadas y ensayadas también pueden usarse para producir un modelo de secuencia-actividad. Típicamente, se usan en un modelo de secuencia-actividad si de hecho están secuenciadas. Cada una de las variantes de proteínas secuenciadas y ensayadas se denomina "observación". Por lo general, cuantas más observaciones se empleen para crear un modelo de secuencia-actividad, mejor será la potencia predictiva de ese modelo de secuencia-actividad.

Hasta la llegada de la tecnología de secuenciación masivamente en paralelo de próxima generación, era difícil secuenciar económicamente más de 10 a 30 péptidos variantes producidos en cualquier ronda de evolución dirigida. Ahora, con la aplicación de la secuenciación de próxima generación, pueden secuenciarse muchas más proteínas variantes producidas en una ronda de evolución dirigida. Como consecuencia, puede usarse un grupo mucho mayor de datos de entrenamiento para producir modelos de secuencia-actividad. Los modelos de secuencia-actividad pueden generarse ahora usando un conjunto de entrenamiento que incluya no sólo los péptidos de mayor rendimiento de una ronda, sino también algunos péptidos que no serían de interés para otras rondas de evolución dirigida, pero cuya información de secuencia-actividad podría aplicarse para producir un modelo de secuencia-actividad más robusto.

En algunas realizaciones, generalmente es deseable producir modelos secuencia-actividad que tengan una buena capacidad para predecir la actividad de una secuencia arbitraria. La potencia predictiva puede caracterizarse por la precisión de la predicción, así como por la consistencia con la que el modelo predice con exactitud la actividad. Además, un modelo puede caracterizarse por su capacidad para predecir con exactitud la actividad en una amplia variedad de espacio de secuencias. Por ejemplo, la potencia predictiva puede caracterizarse en términos de residuos entre las actividades calculadas y reales para un conjunto de péptidos de prueba y/o validación dado. Un modelo con una potencia predictiva generalizada más alta tiende a producir residuos más pequeños y más consistentes a través de diferentes conjuntos de datos de validación. Un modelo que se sobreajusta a un conjunto de datos de prueba tiende a producir residuos más grandes y menos consistentes para los datos de validación, como se muestra en un ejemplo a continuación. Un aspecto de la divulgación proporciona un método para encontrar eficientemente un modelo con una alta potencia predictiva a través de diferentes conjuntos de datos.

A. VISIÓN GENERAL DEL PROCESO DE BÚSQUEDA DE VARIANTES DE PROTEÍNAS MEJORADAS

Los modelos de secuencia-actividad descritos en la presente pueden usarse para ayudar a identificar uno o más "genes" parentales en una biblioteca de variantes inicial para someterlos a evolución dirigida. Después de haber realizado una ronda de evolución, se identifica una nueva biblioteca de variantes, lo que proporciona un nuevo conjunto de observaciones, que luego pueden retroalimentarse como datos para preparar un modelo de secuencia-actividad nuevo o refinado. Este proceso de alternancia entre la preparación de un modelo de secuencia-actividad basado en nuevas observaciones y la realización de una evolución dirigida basada en el modelo de secuencia-actividad puede formar un bucle iterativo de modelado-exploración, que puede repetirse hasta que se obtengan las proteínas y bibliotecas deseadas.

Debido al bucle de retroalimentación entre los modelos de secuencia-actividad y las bibliotecas de variantes, los mejores modelos y las mejores bibliotecas de variantes dependen unos de otros en la exploración de proteínas con actividades mejoradas. Por lo tanto, los cuellos de botella y las mejoras en los dominios de modelado y/o secuenciación pueden afectar a ambos dominios. En algunas realizaciones de la invención, las mejoras de las eficiencias de modelado debidas a mejores técnicas de modelado proporcionan mejores modelos para guiar la exploración de secuencias. En algunas realizaciones, se usan tecnologías de secuenciación de próxima generación para mejorar la velocidad de secuenciación in vitro, así como para proporcionar datos de validación cruzada para mejorar los modelos computacionales in silico.

En algunas realizaciones de la invención, los modelos secuencia-actividad útiles requieren técnicas de modelado matemático robustas y un gran número de "observaciones". Estas observaciones son datos proporcionados en un conjunto de entrenamiento para un modelo. Específicamente, cada observación es un valor de actividad y una secuencia asociada para una variante en una biblioteca. Históricamente, la secuenciación ha sido un paso limitante en el desarrollo de grandes conjuntos de entrenamiento y, en consecuencia, de modelos secuencia-actividad cada vez más robustos. En los métodos que se usan habitualmente en la actualidad, se generan bibliotecas de variantes que pueden tener cientos de variantes. Sin embargo, sólo se secuencia una pequeña fracción de estas variantes. En una ronda típica de evolución dirigida, sólo se secuencian entre aproximadamente 10 y 30 variantes con la actividad más alta. Lo ideal sería secuenciar una fracción mucho mayor de las variantes de la biblioteca, incluyendo algunas variantes con actividades relativamente bajas. Las herramientas de secuenciación de nueva generación han mejorado enormemente la velocidad de secuenciación, haciendo posible incluir las variantes de baja y alta actividad en un conjunto de entrenamiento. En algunas realizaciones, la inclusión de variantes que tienen un intervalo de niveles de actividad da como resultado la producción de modelos que funcionan mejor y/o son mejores para predecir la actividad en un intervalo más amplio de secuencia y espacio de actividad.

Algunos modelos de regresión de secuencia-actividad lineal a los que se hace referencia en la presente incluyen residuos individuales como variables independientes para predecir cualquier actividad de interés. Los modelos de regresión secuencia -actividad lineal no incluyen términos para tener en cuenta de las interacciones entre dos o más residuos. Si una interacción entre dos de los residuos tiene un efecto sinérgico sobre la actividad, un modelo lineal puede proporcionar un valor artificialmente inflado de los coeficientes asociados con los dos residuos que interactúan. Como consecuencia, alguien que trabaje con el modelo puede llegar a la conclusión errónea de que, simplemente haciendo una sustitución de residuos como propone el valor relativamente alto del coeficiente, la actividad de un péptido resultante sería mayor de lo esperado. Esto se debe a que el investigador no comprende que al usar un modelo lineal el aumento de la actividad asociado a la sustitución del residuo es principalmente el resultado de la interacción de esa sustitución con otra sustitución. Si el investigador comprendiera la importancia de esta interacción, entonces podría realizar ambas sustituciones simultáneamente y lograr el aumento de actividad sugerido por el modelo lineal.

Si dos residuos interactúan para suprimir la actividad de manera no lineal, el modelo lineal atribuye a los coeficientes asociados a estos residuos valores inferiores a los que serían apropiados si los residuos se considerasen puramente aislados unos de otros. En otras palabras, realizar una de las sustituciones pero no la otra para los residuos que interactúan producirá un resultado en la actividad superior al que sugeriría el modelo lineal.

Como un modelo lineal puede ser inadecuado cuando las interacciones residuo-residuo tienen un fuerte impacto sobre la actividad, los modelos no lineales con términos de interacción no lineales que tienen en cuenta las interacciones entre residuos son a menudo necesarios para predicciones precisas de la actividad. Sin embargo, los modelos que utilizan términos no lineales plantean retos computacionales y empíricos. En particular, hay que tener en cuenta un gran número de posibles términos de interacción a la hora de desarrollar/utilizar un modelo, lo que requiere una cantidad considerable de cálculos. Una limitación mucho mayor es el número potencial de observaciones necesarias para producir un modelo con un número significativo de términos de interacción residuo-residuo. Además, puede haber una tendencia a que la técnica de creación de modelos se ajuste en exceso a los datos, dado un número particular de observaciones disponibles. Para abordar este reto, es una consideración importante en el desarrollo de muchos modelos seleccionar y limitar cuidadosamente los términos de interacción proporcionados en el modelo secuencia-actividad.

La Figura 1 presenta un diagrama de flujo que muestra una implementación de un proceso de preparación de un modelo de secuencia-actividad. Como se representa, un proceso 100 comienza en un bloque 103 para proporcionar datos de secuencia y actividad para genes variantes ("observaciones"). Los datos de secuencia pueden tomarse, por ejemplo, de un conjunto de entrenamiento que comprende secuencias de residuos para una biblioteca de variantes de proteínas inicial o mejorada. Típicamente, estos datos incluyen información completa o parcial de la secuencia de residuos, junto con un valor de actividad para cada proteína de la biblioteca. En algunos casos, se proporcionan juntos múltiples tipos de actividades (por ejemplo, datos de constante de velocidad y datos de estabilidad térmica) en el conjunto de entrenamiento. También pueden considerarse otras fuentes de datos, según lo determinen los resultados deseados. Algunas fuentes de datos adecuadas incluyen, pero no se limitan a, referencias bibliográficas que describen información sobre péptidos particulares de relevancia para el modelo de actividad de secuencia en construcción. Fuentes de información adicionales incluyen, entre otras, rondas anteriores o diferentes de evolución dirigida en el mismo proyecto. De hecho, se pretende que la información derivada de rondas anteriores de evolución dirigida (usando cualquier método adecuado, incluyendo pero no limitándose a los proporcionados en la presente) se use en el desarrollo de bibliotecas, variantes, etc. producidas posteriormente.

En muchas realizaciones, los miembros individuales de la biblioteca de variantes de proteínas representan una amplia variedad de secuencias y actividades. Esto facilita la generación de un modelo secuencia-actividad que sea aplicable en una amplia región del espacio de secuencias. Las técnicas para generar tales bibliotecas diversas incluyen, pero no se limitan a, la variación sistemática de secuencias de proteínas y técnicas de evolución dirigida, como se describe en la presente. Sin embargo, en algunas realizaciones alternativas, es deseable generar modelos a partir de secuencias de genes en una familia de genes particular (por ejemplo, una quinasa particular que se encuentra en múltiples especies u organismos). Como en todos los miembros de la familia muchos residuos serán idénticos, el modelo describe sólo aquellos residuos que varían. Por tanto, en algunas realizaciones, los modelos estadísticos basados en tales conjuntos de entrenamiento relativamente pequeños, comparados con el conjunto de todas las variantes posibles, son válidos en un sentido local. Es decir, los modelos son válidos sólo para las observaciones dadas de las variantes dadas. En algunas realizaciones, el objetivo no es encontrar una función de ajuste global, ya que se reconoce que en algunos modelos, esto está más allá de la capacidad y/o necesidad del sistema o sistemas del modelo en consideración.

Los datos de actividad pueden obtenerse usando cualquier medio adecuado conocido en la técnica, incluyendo pero no limitado a ensayos y/o cribas adecuadamente diseñados para medir magnitudes de la actividad/actividades de interés. Tales técnicas son bien conocidas y no son esenciales para la presente invención. Los principios para el diseño de ensayos o cribas apropiados son ampliamente comprendidos y conocidos en la técnica. Las técnicas para obtener secuencias de proteínas también son bien conocidas y no son esenciales para la presente invención. Como se ha mencionado, pueden usarse tecnologías de secuenciación de próxima generación. La actividad usada en las realizaciones descritas en la presente puede ser la estabilidad de la proteína (por ejemplo, la estabilidad térmica). Sin embargo, muchas realizaciones importantes consideran otras actividades como la actividad catalítica, la resistencia a patógenos y/o toxinas, la actividad terapéutica, la toxicidad y similares. De hecho, no se pretende que la presente invención se limite a ningún método de ensayo/cribado y/o método de secuenciación en particular, ya que en la presente invención encuentra uso cualquier método adecuado conocido en la técnica.

Después de haber generado o adquirido los datos del conjunto de entrenamiento, el proceso los usa para generar un modelo base de secuencia-actividad que predice la actividad en función de la información de la secuencia. Ver el bloque 105. Este modelo es una expresión, algoritmo u otra herramienta que predice la actividad relativa de una proteína en particular cuando se le proporciona información de la secuencia de dicha proteína. En otras palabras, se introduce la información de la secuencia de la proteína y se obtiene una predicción de la actividad. En algunas realizaciones, el modelo base no incluye ningún término de interacción. En tales casos, el modelo base puede describirse como un "modelo lineal". En otras realizaciones, el modelo base incluye todos los términos de interacción disponibles, en cuyo caso el modelo base puede describirse como un modelo no lineal o un modelo de interacción.

Para muchas realizaciones, el modelo base puede clasificar la contribución de varios residuos a la actividad. A continuación se analizan los métodos para generar dichos modelos, que se engloban en el ámbito del aprendizaje automático (por ejemplo, regresión de mínimos cuadrados parciales (PLS), regresión de componentes principales

(PCR), regresión lineal múltiple (MLR), regresión lineal bayesiana), junto con el formato de las variables independientes (información de secuencia), el formato de la variable o variables dependientes (actividad) y la forma del propio modelo (por ejemplo, una expresión lineal de primer orden).

Después de que se haya generado un modelo base de actividad secuencial, el proceso añade o sustrae iterativamente términos de interacción de un grupo de términos de interacción disponibles hacia o desde el modelo base y evalúa los nuevos modelos resultantes para su mejora sobre el modelo base para producir un modelo final. Ver el bloque 107. Cuando el modelo base incluye todos los términos de interacción disponibles, el proceso sustrae dichos términos de manera escalonada. Cuando el modelo base no incluye términos de interacción, el proceso añade dichos términos de manera escalonada.

Al evaluar un nuevo modelo, los métodos de la presente divulgación no sólo tienen en cuenta la varianza que un modelo tiene en cuenta dado un conjunto de datos, sino también la capacidad del modelo para predecir nuevos datos. En algunas realizaciones, este enfoque de selección de modelos penaliza los modelos que tienen más coeficientes/parámetros que los modelos equivalentes que tienen menos coeficientes/parámetros para evitar el ajuste excesivo del modelo al conjunto de datos dado. Ejemplos de métodos de selección incluyen, pero no se limitan a, el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC), y variaciones de los mismos.

En una serie de modelos anidados, como en los modelos de regresión con un número progresivamente mayor de términos de interacción (y coeficientes asociados) que un modelo base, los modelos más complejos proporcionan ajustes igual de buenos o mejores que los más sencillos, incluso si los coeficientes adicionales son espurios, porque el modelo más complejo disfruta de grados de libertad adicionales. Ciertas realizaciones de la presente divulgación emplean métodos de selección de modelos que penalizan los modelos más complejos en la medida en que la ganancia en la bondad del ajuste es más que compensada por el coste de los parámetros espurios.

A continuación se presentan algoritmos ejemplares para generar modelos de secuencia-actividad de acuerdo con las operaciones de los bloques 105 y 107. Tales técnicas incluyen, pero no se limitan a, técnicas paso a paso que evitan la inclusión de términos de interacción adicionales en un modelo. Sin embargo, no se pretende que la presente divulgación se limite a estos ejemplos específicos.

En un aspecto, la presente divulgación proporciona métodos para preparar un modelo de secuencia-actividad que puede ayudar a identificar moléculas biológicas para afectar a una actividad deseada. En algunas realizaciones, como se define adicionalmente en las realizaciones, el método comprende: (a) recibir datos de secuencia y actividad para una pluralidad de moléculas biológicas; (b) preparar un modelo base a partir de los datos de secuencia y actividad, en donde el modelo base predice la actividad como una función de la presencia o ausencia de subunidades de la secuencia; (c) preparar por lo menos un nuevo modelo añadiendo o sustrayendo por lo menos un nuevo término de interacción a o desde el modelo base, en donde el nuevo término de interacción representa la interacción entre dos o más subunidades que interactúan; (d) determinar la capacidad del por lo menos un nuevo modelo para predecir la actividad en función de la presencia o ausencia de las subunidades; y (e) determinar si añadir o sustraer el nuevo término de interacción a o del modelo base basándose en la capacidad del por lo menos un nuevo modelo para predecir la actividad determinada en (d) y con un sesgo en contra de añadir el nuevo término de interacción. El modelo derivado puede usarse entonces en varias aplicaciones, como en la evolución dirigida de bibliotecas de proteínas para identificar proteínas con las actividades y propiedades biológicas deseadas.

En donde el método determina que el nuevo término de interacción debe añadirse al modelo base para producir un modelo actualizado, el método también incluye: (f) repetir (c) usando el modelo actualizado en lugar del modelo base y añadir o sustraer un término de interacción diferente del añadido/sustraído en (c); y (g) repetir (d) y (e) usando el modelo actualizado en lugar del modelo base. En algunas realizaciones, el método incluye además: (h) repetir (f) y (g) usar un modelo actualizado adicional.

Después de que se hayan seleccionado las observaciones para un conjunto de entrenamiento y se haya elegido una técnica matemática para producir el modelo secuencia-actividad, se crea el modelo base. El modelo base se genera típicamente sin tener en cuenta su capacidad de predicción. Simplemente se produce de acuerdo con un procedimiento definido para producir un modelo base a partir de las observaciones disponibles (es decir, el conjunto de observaciones), como se describe en la presente. Como se ha indicado anteriormente, los modelos de secuencia pueden describir varias secuencias, mientras que en algunas realizaciones, los modelos describen proteínas. En este último caso, el modelo base es simplemente un modelo lineal con un único término para cada una de las mutaciones presentes en la colección de péptidos usada para crear el conjunto de entrenamiento. En estas realizaciones, el modelo base no incluye ningún término que represente interacciones entre residuos en los péptidos. En algunas realizaciones, el modelo base no incluye un término separado para todas y cada una de las mutaciones presentes en el conjunto de observación.

En enfoques alternativos, el modelo base incluye no sólo los términos que describen cada una de las mutaciones de manera aislada, sino que además incluye términos para todos los posibles residuos que interactúan. En el caso extremo, en el modelo base se usan todas las interacciones imaginables entre las mutaciones señaladas.

Esto incluye un término para cada una de las interacciones por pares entre las mutaciones, así como términos para cada una de las posibles interacciones de tres residuos, así como cada una de las posibles interacciones de cuatro residuos, etc. Algunas realizaciones incluyen sólo las interacciones por pares o las interacciones por pares y las interacciones de tres vías. Una interacción de tres vías es una interacción que afecta a la actividad entre tres subunidades distintas.

En una o más realizaciones que usan un modelo lineal simple como modelo base, los esfuerzos posteriores para mejorar el modelo incluyen la adición de nuevos términos que representan interacciones distintas. En realizaciones alternativas en las que el modelo base incluye todos los términos lineales y no lineales, los esfuerzos posteriores para mejorar el modelo implican la eliminación selectiva de algunos de los términos de interacción no lineales.

En una o más realizaciones de la invención, el proceso de mejora del modelo base implica añadir o sustraer iterativamente términos de interacción del modelo base para determinar si el modelo resultante mejora suficientemente la calidad del modelo. En cada iteración, se determina la potencia predictiva del modelo actual y se compara con otro modelo, por ejemplo, el modelo base o el modelo actualizado.

En las realizaciones en las que una medida de potencia predictiva ya tiene en cuenta la capacidad de un modelo para generalizarse a otros conjuntos de datos, esa medida por sí sola puede determinar si debe seleccionarse un modelo candidato. Por ejemplo, una medida como AIC o BIC tiene en cuenta tanto la verosimilitud del modelo (o error residual) como el número de parámetros. Una "función de verosimilitud" o "verosimilitud" de un modelo es una función de los parámetros de un modelo estadístico. La verosimilitud de un conjunto de valores de parámetros dados unos resultados observados es igual a la probabilidad de esos resultados observados dados esos valores de parámetros, es decir, $L(\theta|x) = P(x|\theta)$. A continuación se describe un cálculo ejemplar de la verosimilitud de un modelo. Medidas como AIC y BIC están sesgadas en contra de un modelo que tiene más parámetros si el modelo con más parámetros captura la misma cantidad de varianza de datos que un modelo que tiene menos parámetros. Si una medida de potencia predictiva sólo tiene en cuenta el error residual, debe considerarse la magnitud de la mejora en el error residual para determinar si incorporar o no el cambio asociado con la iteración actual en el mejor modelo actualizado. Esto puede lograrse comparando la magnitud de la mejora con un umbral. Si la magnitud es menor que el umbral, no se acepta el cambio considerado en la iteración actual. Si, por el contrario, la magnitud de la mejora supera el umbral, entonces el cambio considerado se incorpora al modelo actualizado y el modelo actualizado sirve como nuevo mejor modelo para las iteraciones restantes.

En ciertas realizaciones, cada iteración considera la adición o sustracción de un único término de interacción del mejor modelo actual en consideración. En el caso de un modelo aditivo, es decir, el caso para el que el modelo base contiene sólo términos lineales, puede considerarse un conjunto de todos los términos de interacción disponibles. Se considera sucesivamente cada uno de estos términos de interacción hasta que se completa el proceso y se obtiene un mejor modelo final.

En algunos casos, tras determinar que el proceso ha convergido efectivamente y que es improbable que se produzcan nuevas mejoras, el proceso de generación del modelo finaliza antes de que se hayan considerado todos los términos de interacción disponibles en el grupo.

La Figura 2 ilustra cómo puede usarse iterativamente un modelo para guiar la creación de nuevas bibliotecas de variantes de proteínas con el propósito de explorar el espacio de secuencia y actividad de proteínas, en un proceso (Ver, 200). Después de que se haya generado un modelo final, el modelo final se emplea para identificar múltiples posiciones de residuos (por ejemplo, la posición 35) o valores de residuos específicos (por ejemplo, glutamina en la posición 35) que se predice que impactan a la actividad. Ver el bloque 207. Además de identificar tales posiciones, el modelo puede usarse para "clasificar" las posiciones de residuos o valores de residuos basándose en sus contribuciones a la actividad deseada (¿actividades?). Por ejemplo, el modelo puede predecir que la glutamina en la posición 35 tiene el efecto positivo más pronunciado sobre la actividad; la fenilalanina en la posición 208 tiene el segundo efecto positivo más pronunciado sobre la actividad; y así sucesivamente. En un enfoque específico descrito a continuación, se emplean coeficientes de regresión PLS o PCR para clasificar la importancia de residuos específicos. En otro enfoque específico, se emplea una matriz de carga PLS para clasificar la importancia de posiciones específicas de residuos.

Después de que el proceso haya identificado los residuos que afectan a la actividad, se seleccionan algunos de ellos para su variación, como se indica en el bloque 209 (Figura 2). Esto se hace con el propósito de explorar el espacio de secuencias. Los residuos se seleccionan usando cualquiera de una serie de diferentes protocolos de selección, algunos de los cuales se describen a continuación. En un ejemplo ilustrativo, se conservan (es decir, no se varían) los residuos específicos que se prevé tendrán el impacto más beneficioso sobre la actividad. Sin embargo, se seleccionan para la variación un cierto número de otros residuos que se prevé que tengan un impacto menor. En otro ejemplo ilustrativo, se seleccionan para la variación las posiciones de los residuos que tienen el mayor impacto en la actividad, pero sólo si se encuentra que varían en los miembros de alto rendimiento del conjunto de entrenamiento. Por ejemplo, si el modelo predice que la posición del residuo 197 tiene el mayor impacto sobre la actividad, pero todas

o la mayoría de las proteínas con alta actividad tienen leucina en esta posición, en este enfoque no se seleccionaría para la variación la posición 197. En otras palabras, todas o la mayoría de las proteínas de una biblioteca de próxima generación tendrían leucina en la posición 197. Sin embargo, si algunas proteínas "buenas" tuvieran valina en esta posición y otras tuvieran leucina, entonces el proceso elegiría variar el aminoácido en esta posición. En algunos casos, se descubrirá que una combinación de dos o más residuos que interactúan tiene el mayor impacto sobre la actividad. Por lo tanto, en algunas estrategias, estos residuos se covarian.

Después de haber identificado los residuos para la variación, el método genera a continuación una nueva biblioteca de variantes que tiene la variación de residuos especificada. Ver el bloque 211 (Figura 2). Existen varias metodologías disponibles para este propósito. En un ejemplo, se realiza un mecanismo de generación de diversidad basado en recombinación in vitro o in vivo para generar la nueva biblioteca de variantes. Tales procedimientos pueden emplear oligonucleótidos que contienen secuencias o subsecuencias para codificar las proteínas de la biblioteca de variantes parental. Algunos de los oligonucleótidos estarán estrechamente relacionados, diferenciándose sólo en la elección de codones para aminoácidos alternativos seleccionados para la variación en 209. El mecanismo de generación de diversidad basado en la recombinación puede realizarse durante uno o varios ciclos. Si se usan múltiples ciclos, cada uno implica un paso de cribado para identificar qué variantes tienen un rendimiento aceptable para ser usadas en un ciclo de recombinación posterior. Se trata de una forma de evolución dirigida. Sin embargo, no se pretende que la presente invención se limite a ningún método específico de generación de diversidad basado en la recombinación, ya que cualquier método/técnica adecuado encuentra uso en la presente invención.

En un ejemplo ilustrativo adicional, se elige una secuencia de proteína de "referencia" y los residuos seleccionados en 209 de la Figura 2 se "alternan" para identificar miembros individuales de la biblioteca de variantes. Las nuevas proteínas identificadas de este modo se sintetizan mediante una técnica apropiada para generar la nueva biblioteca. En un ejemplo, la secuencia de referencia puede ser un miembro de alto rendimiento del conjunto de entrenamiento o una "mejor" secuencia predicha por un modelo PLS o PCR.

En otro ejemplo ilustrativo, los residuos para la variación en una ronda de evolución dirigida se seleccionan en una única secuencia parental. El progenitor puede identificarse usando resultados modelo de una ronda anterior de evolución dirigida o usando datos que identifiquen el miembro de la biblioteca que tenga el mejor rendimiento en el ensayo. Los oligonucleótidos para la siguiente ronda de evolución dirigida pueden definirse para que incluyan partes de la estructura principal del progenitor seleccionado con una o más mutaciones predichas algorítmicamente a partir de un modelo de actividad de secuencia para la ronda actual. Estos oligonucleótidos pueden producirse usando cualquier medio adecuado, incluyendo, entre otros, métodos sintéticos.

Una vez producida la nueva biblioteca, se comprueba su actividad, como se indica en el bloque 213 (Figura 2). Idealmente, la nueva biblioteca proporciona uno o más miembros con mejor actividad que la observada en la biblioteca anterior. Sin embargo, incluso sin dicha ventaja, la nueva biblioteca puede proporcionar información beneficiosa. Sus miembros pueden emplearse para generar modelos mejorados que tengan en cuenta los efectos de las variaciones seleccionadas en 209 (Figura 2), y de este modo predecir con mayor precisión la actividad en regiones más amplias del espacio de secuencias. Además, la biblioteca puede representar un pasaje en el espacio de secuencia desde un máximo local hacia un máximo global (por ejemplo, en actividad).

Dependiendo del objetivo del proceso 200 (Figura 2), en algunas realizaciones, es deseable generar una serie de nuevas bibliotecas de variantes de proteínas, cada una de las cuales proporciona nuevos miembros de un conjunto de entrenamiento. El conjunto de entrenamiento actualizado se usa luego para generar un modelo mejorado. Para lograr el modelo mejorado, el proceso 200 se muestra con una operación de decisión como se muestra en el bloque 215, que determina si se debe producir otra biblioteca de variantes de proteínas. Pueden usarse varios criterios para tomar esta decisión. Ejemplos de criterios de decisión incluyen pero no se limitan al número de bibliotecas de variantes de proteínas generadas hasta el momento, la actividad de las proteínas principales de la biblioteca actual, la magnitud de la actividad deseada y el nivel de mejora observado en las nuevas bibliotecas recientes.

Suponiendo que el proceso se use para continuar con una nueva biblioteca, el proceso vuelve a la operación del bloque 100 (Figura 2) donde se genera un nuevo modelo de secuencia-actividad a partir de los datos de secuencia y actividad obtenidos para la biblioteca de variantes de proteínas actual. En otras palabras, los datos de secuencia y actividad para la biblioteca de variantes de proteínas actual sirven como parte del conjunto de entrenamiento para el nuevo modelo (o pueden servir como todo el conjunto de entrenamiento). A partir de entonces, las operaciones mostradas en los bloques 207, 209, 211, 213, y 215 (Figura 2) se realizan como se ha descrito anteriormente, pero con el nuevo modelo.

Cuando se determina que se ha alcanzado el punto final del método, se termina el ciclo ilustrado en la Figura 2 y no se genera ninguna biblioteca nueva. En ese momento, el proceso simplemente se termina o, en algunas realizaciones, se seleccionan una o más secuencias de una o más de las bibliotecas para su desarrollo y/o fabricación. Ver el bloque 217.

B. GENERACIÓN DE OBSERVACIONES

Las bibliotecas de variantes de proteínas son grupos de múltiples proteínas que tienen uno o más residuos que varían de un miembro a otro en una biblioteca. Estas bibliotecas pueden generarse usando los métodos descritos en la presente y/o cualquier otro medio adecuado conocido en la técnica. Estas bibliotecas encuentran uso para proporcionar datos para conjuntos de entrenamiento usados para generar modelos de secuencia-actividad de acuerdo con varias realizaciones de la presente invención. El número de proteínas incluidas en una biblioteca de variantes de proteínas depende a menudo de la aplicación y del coste asociado a su generación. No se pretende que la presente invención se limite a un número concreto de proteínas en las bibliotecas de proteínas usadas en los métodos de la presente invención. Tampoco se pretende que la presente invención se limite a ninguna biblioteca o bibliotecas de variantes de proteínas en particular.

En un ejemplo, la biblioteca de variantes de proteínas se genera a partir de una o más proteínas de origen natural, que pueden estar codificadas por una única familia de genes. Pueden usarse otros puntos de partida que incluyen pero no se limitan a recombinantes de proteínas conocidas o nuevas proteínas sintéticas. A partir de estas proteínas semilla o de partida, la biblioteca puede generarse mediante varias técnicas. En un caso, la biblioteca se genera mediante recombinación mediada por fragmentación del ADN como se describe en Stemmer (1994) Proceedings of the National Academy of Sciences, USA, 10747-10751 y la WO 95/22625, recombinación mediada por oligonucleótidos sintéticos como se describe en Ness et al. (2002) Nature Biotechnology 20:1251-1255 y la WO 00/42561, o ácidos nucleicos que codifican parte o la totalidad de una o más proteínas parentales. También pueden usarse combinaciones de estos métodos (por ejemplo, recombinación de fragmentos de ADN y oligonucleótidos sintéticos), así como otros métodos basados en recombinación descritos en, por ejemplo, la WO97/20078 y la WO98/27230. En la presente invención encuentra uso cualquier método adecuado usado para generar bibliotecas de variantes de proteínas. De hecho, no se pretende que para producir bibliotecas de variantes la presente invención se limite a ningún método particular.

En algunas realizaciones, puede emplearse una única secuencia "de partida" (que puede ser una secuencia "antepasada") para definir un grupo de mutaciones usadas en el proceso de modelado. En algunas realizaciones, por lo menos una de las secuencias de partida es una secuencia de tipo salvaje.

En ciertas realizaciones, las mutaciones (a) se identifican en la bibliografía por afectar a la especificidad, la selectividad, la estabilidad u otra propiedad beneficiosa del sustrato y/o (b) se predicen computacionalmente para mejorar los patrones de plegamiento de la proteína (por ejemplo, el empaquetamiento de los residuos interiores de una proteína), la unión de ligandos, las interacciones de subunidades, la transposición de familias entre múltiples homólogos diversos, etc. Como alternativa, las mutaciones pueden introducirse físicamente en la secuencia de partida y cribarse los productos de expresión para determinar las propiedades beneficiosas. La mutagénesis dirigida al sitio es un ejemplo de técnica útil para introducir mutaciones, aunque puede usarse cualquier método adecuado. Por tanto, alternativa o adicionalmente, los mutantes pueden obtenerse mediante síntesis genética, mutagénesis aleatoria saturante, bibliotecas combinatorias semisintéticas de residuos, evolución dirigida, recombinación recursiva de secuencias ("RSR") (ver, por ejemplo, la Solicitud de Patente de Estados Unidos Nº 2006/0223143), trasposición de genes, PCR propensa a errores, y/o cualquier otro método adecuado. Un ejemplo de procedimiento de mutagénesis de saturación adecuado se describe en la Solicitud de Patente Publicada de Estados Unidos Nº 20100093560.

No es necesario que la secuencia de partida sea idéntica a la secuencia de aminoácidos de la proteína salvaje. Sin embargo, en algunas realizaciones, la secuencia de partida es la secuencia de la proteína salvaje. En algunas realizaciones, la secuencia de partida incluye mutaciones no presentes en la proteína salvaje. En algunas realizaciones, la secuencia de partida es una secuencia de consenso derivada de un grupo de proteínas que tienen una propiedad común, por ejemplo, una familia de proteínas.

Una lista representativa no limitativa de familias o clases de enzimas que pueden servir como fuentes de secuencias parentales incluye, entre otras, las siguientes: oxidorreductasas (C.E.1); transferasas (C.E.2); hidroliasas (C.E.3); liasas (C.E.4); isomerasas (C.E.5) y ligasas (C.E.6). Los subgrupos más específicos pero no limitativos de oxidorreductasas incluyen las deshidrogenasas (por ejemplo, alcohol deshidrogenasas (carbonil reductasas), xilulosa reductasas, aldehído reductasas, farnesol deshidrogenasa, lactato deshidrogenasas, arabinosa deshidrogenasas, glucosa deshidrogenasa, fructosa deshidrogenasas, xilosa reductasas y succinato deshidrogenasas), oxidasas (por ejemplo, glucosa oxidasas, hexosas oxidasas, galactosa oxidasas y lacasas), monoamino oxidasas, lipoxigenasas, peroxidasas, aldehído deshidrogenasas, reductasas, acil-[acil-portador-proteína] reductasas de cadena larga, acil-CoA deshidrogenasas, ene-reductasas, sintasas (por ejemplo, glutamato sintasas), nitrato reductasas, mono y di-oxigenasas y catalasas. Subgrupos más específicos pero no limitativos de transferasas incluyen metil, amidino y carboxil transferasas, transcetolasas, transaldolasas, aciltransferasas, glicosiltransferasas, transaminasas, transglutaminasas y polimerasas. Los subgrupos más específicos pero no limitativos de hidrolasas incluyen las éster hidrolasas, peptidasas, glicosilasas, amilasas, celulasas, hemicelulasas, xilanasas, quitinasas, glucosidasas, glucanasas, glucoamilasas, acilasas, galactosidasas, pullulanasas, fitasas, lactasas, arabinosidasas, nucleosidasas, nitrilasas, fosfatasas, lipasas, fosfolipasas, proteasas, ATPasas y deshalogenasas. Subgrupos más específicos pero no limitativos de liasas incluyen descarboxilasas, aldolasas, hidrasas, deshidratasas (por ejemplo, anhidrasas carbónicas), sintasas (por ejemplo, isopreno, pineno y farneseno sintasas), pectinasas (por ejemplo, pectina liasas) y

halohidrina deshidrogenasas. Subgrupos más específicos, pero no limitativos, de isomerasas incluyen racemasas, epimerasas, isomerasas (por ejemplo, isomerasas de xilosa, arabinosa, ribosa, glucosa, galactosa y manosa), tautomerasas y mutasas (por ejemplo, mutasas de transferencia de acilo, fosfomutasas y aminomutasas). Los subgrupos más específicos pero no limitativos de ligasas incluyen las estero sintasas. Otras familias o clases de enzimas que pueden usarse como fuentes de secuencias parentales incluyen transaminasas, proteasas, quinasas y sintasas. Esta lista, aunque ilustra ciertos aspectos específicos de las posibles enzimas de la divulgación, no se considera exhaustiva y no retrata las limitaciones ni circunscribe el alcance de la divulgación.

En algunos casos, las enzimas candidatas útiles en los métodos descritos en la presente son capaces de catalizar una reacción enantioselectiva, como una reacción de reducción enantioselectiva, por ejemplo. Tales enzimas pueden usarse para elaborar productos intermedios útiles en la síntesis de compuestos farmacéuticos, por ejemplo.

En algunas realizaciones, las enzimas candidatas se seleccionan entre endoxilanasas (EC 3.2.1.8); β-xilosidasas (EC 3.2.1.37); alfa-L-arabinofuranosidasas (EC 3.2.1.55); alfa-glucuronidasas (EC 3.2.1.139); acetilxilanesterasas (EC 3.1.1.72); feruloil esterasas (EC 3.1.1.73); cumaroil esterasas (EC 3.1.1.73); alfa-galactosidasas (EC 3.2.1.22); beta-galactosidasas (EC 3.2.1.23); beta-mananasas (EC 3.2.1.78); beta-mannosidasas (EC 3.2.1.25); endo-poligalacturonasas (EC 3.2.1.15); pectina metil esterasas (EC 3.1.1.11); endo-galactanasas (EC 3.2.1.89); pectina acetil esterasas (EC 3.1.1.6); endo-pectina liasas (EC 4.2.2.10); pectato liasas (EC 4.2.2.2); alfa ramnosidasas (EC 3.2.1.40); exo-poli-alfa-galacturonosidasa (EC 3.2.1.82); 1,4-alfa-galacturonidasa (EC 3.2.1.67); exopoligalacturonato liasas (EC 4.2.2.9); ramnogalacturonano endilasas EC (4.2.2.B3); ramnogalacturonano acetilesterasas (EC 3..2.1.B11); ramnogalacturonano galacturonohidrolasas (EC 3.2.1.B11); endo-arabinanasas (EC 3.2.1.99); lacasas (EC 1.10.3.2); peroxidasas dependientes del manganeso (EC 1.10.3.2); amilasas (EC 3.2.1.1), glucoamilasas (EC 3.2.1.3), proteasas, lipasas y lignina peroxidasas (EC 1.11.1.14). En las composiciones de la presente invención puede usarse cualquier combinación de una, dos, tres, cuatro, cinco o más de cinco enzimas.

En una o más realizaciones de la invención, para generar la biblioteca se modifica una única secuencia de partida de varias maneras. En algunas realizaciones, la biblioteca se genera variando sistemáticamente los residuos individuales de la secuencia de partida. En un ejemplo ilustrativo, se emplea una metodología de diseño de experimentos (DOE) para identificar las secuencias variadas sistemáticamente. En otro ejemplo, se usa un procedimiento de "laboratorio húmedo", como la recombinación mediada por oligonucleótidos, para introducir cierto nivel de variación sistemática. No se pretende que la presente invención se limite a ningún método en particular para generar secuencias sistemáticamente variadas, ya que puede usarse cualquier método adecuado.

Como se usa en la presente, el término "secuencias sistemáticamente variadas" se refiere a un conjunto de secuencias en las que cada residuo se ve en múltiples contextos. En principio, el nivel de variación sistemática puede cuantificarse por el grado en que las secuencias son ortogonales entre sí (es decir, máximamente diferentes en comparación con la media). En algunas realizaciones, el proceso no depende de que las secuencias sean máximamente ortogonales. Sin embargo, la calidad del modelo mejorará en relación directa con la ortogonalidad del espacio de secuencias probado. En un ejemplo ilustrativo sencillo, se varía sistemáticamente una secuencia peptídica identificando dos posiciones de residuos, cada una de las cuales puede tener uno de dos aminoácidos diferentes. Una biblioteca de máxima diversidad incluye las cuatro secuencias posibles. Esta variación sistemática máxima aumenta exponencialmente con el número de posiciones variables; por ejemplo, en 2^N, cuando hay 2 opciones en cada una de las N posiciones de residuos. Sin embargo, los expertos en la técnica reconocerán fácilmente que no es necesaria una variación sistemática máxima. La variación sistemática proporciona un mecanismo para identificar un conjunto relativamente pequeño de secuencias para las pruebas que proporciona un buen muestreo del espacio de secuencias.

Las variantes de proteínas con secuencias sistemáticamente variadas pueden obtenerse de varias maneras mediante técnicas bien conocidas por los expertos en la técnica. Como se ha indicado, los métodos adecuados incluyen, entre otros, métodos basados en la recombinación que generan variantes basadas en una o más secuencias de polinucleótidos "parentales". Las secuencias de polinucleótidos pueden recombinarse usando una variedad de técnicas, incluyendo, por ejemplo, la digestión con ADNasa de los polinucleótidos que van a recombinarse, seguido de la ligadura y/o el reensamblaje por PCR de los ácidos nucleicos. Estos métodos incluyen, entre otros, los descritos en, por ejemplo, Stemmer (1994) Proceedings of the National Academy of Sciences USA, 91:10747-10751, Patente de Estados Unidos Nº 5,605,793, "Methods for In Vitro Recombination," Patente de Estados Unidos Nº 5,811,238, "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination", Patente de Estados Unidos Nº 5,830,721, "DNA Mutagenesis by Random Fragmentation and Reassembly", Patente de Estados Unidos Nº 5,834,252, "End Complementary Polymerase Reaction", Patente de Estados Unidos Nº 5,837,458, "Methods and Compositions for Cellular and Metabolic Engineering", WO98/42832, "Recombination of Polynucleotide Sequences Using Random or Defined Primers", WO 98/27230, "Methods and Compositions for Polypeptide Engineering", WO 99/29902, "Method for Creating Polynucleotide and Polypeptide Sequences" y similares.

Para generar bibliotecas de variantes de proteínas con variación sistemática también son especialmente adecuados los métodos de recombinación sintética. En los métodos de recombinación sintética, se sintetiza una pluralidad de oligonucleótidos que codifican colectivamente una pluralidad de los genes que se van a recombinar. En

algunas realizaciones, los oligonucleótidos codifican colectivamente secuencias derivadas de genes parentales homólogos. Por ejemplo, los genes homólogos de interés se alinean usando un programa de alineación de secuencias como BLAST (consultar, por ejemplo, Atschul, et al., Journal of Molecular Biology, 215:403-410 (1990). Se anotan los nucleótidos correspondientes a las variaciones de aminoácidos entre los homólogos. Estas variaciones se restringen opcionalmente a un subconjunto del total de variaciones posibles basándose en el análisis de covariación de las secuencias parentales, la información funcional de las secuencias parentales, la selección de cambios conservadores o no conservadores entre las secuencias parentales u otros criterios adecuados. Las variaciones se incrementan opcionalmente para codificar una diversidad adicional de aminoácidos en posiciones identificadas, por ejemplo, mediante el análisis de covariación de las secuencias parentales, la información funcional de las secuencias parentales, la selección de cambios conservadores o no conservadores entre las secuencias parentales o la tolerancia aparente de una posición a la variación. El resultado es una secuencia génica degenerada que codifica una secuencia de aminoácidos de consenso derivada de las secuencias génicas parentales, con nucleótidos degenerados en posiciones que codifican variaciones de aminoácidos. Se diseñan oligonucleótidos que contienen los nucleótidos necesarios para ensamblar la diversidad presente en el gen degenerado. Los detalles referentes a tales enfoques pueden encontrarse, por ejemplo, en Ness et al. (2002), Nature Biotechnology, 20:1251-1255, WO 00/42561, "Oligonucleotide Mediated Nucleic Acid Recombination", WO 00/42560, "Methods for Making Character Strings, Polynucleotides and Polypeptides having Desired Characteristics", WO 01/75767, "In Silico Cross-Over Site Selection", y WO 01/64864, "Single-Stranded Nucleic Acid Template-Mediated Recombination and Nucleic Acid Fragment Isolation". Las secuencias variantes de polinucleótidos identificadas pueden transcribirse y traducirse, ya sea in vitro o in vivo, para crear un conjunto o biblioteca de secuencias variantes de proteínas.

El conjunto de secuencias sistemáticamente variadas también puede diseñarse a priori usando métodos de diseño de experimentos (DOE) para definir las secuencias en el conjunto de datos. Una descripción de los métodos DOE puede encontrarse en Diamond, W. J. (2001) Practical Experiment Designs: for Engineers and Scientists, John Wiley & Sons y en "Practical Experimental Design for Engineers and Scientists" de William J Drummond (1981) Van Nostrand Reinhold Co New York, "Statistics for experimenters" George E. P. Box, William G Hunter y J. Stuart Hunter (1978) John Wiley and Sons, New York, o, por ejemplo en la web en itl.nist.gov/div898/handbook/. Hay varios paquetes informáticos para realizar las operaciones matemáticas relevantes, como Statistics Toolbox (MATLAB®), JMP®, STATISTICA® y STAT-EASE® DESIGN EXPERT®. El resultado es un conjunto de datos sistemáticamente variado y ortogonalmente disperso de secuencias que es adecuado para construir el modelo secuencia-actividad de la presente invención. Los conjuntos de datos basados en DOE también pueden generarse fácilmente usando diseños Plackett-Burman o Factorial Fraccional, como se conoce en la técnica. Diamond, W. J. (2001).

En ingeniería y ciencias químicas, los diseños factoriales fraccionados se usan para definir menos experimentos en comparación con los diseños factoriales completos. En estos métodos, un factor varía (es decir, se "alterna") entre dos o más niveles. Se usan técnicas de optimización para garantizar que los experimentos elegidos sean lo más informativos posible a la hora de tener en cuenta la varianza del espacio factorial. Los mismos enfoques de diseño (por ejemplo, factorial fraccional, diseño D-óptimo) pueden aplicarse en manipulación de proteínas para construir menos secuencias en las que un número determinado de posiciones se alternan entre dos o más residuos. En algunas realizaciones, este conjunto de secuencias proporciona una descripción óptima de la varianza sistemática presente en el espacio de secuencias de la proteína en cuestión.

Un ejemplo ilustrativo del enfoque DOE aplicado a la ingeniería de proteínas incluye las siguientes operaciones:

1) Identificar las posiciones a alternar basándose en los principios descritos en la presente (por ejemplo, presentes en las secuencias parentales, nivel de conservación, etc.).
2) Crear un experimento DOE usando uno de los paquetes de software estadístico disponibles habitualmente, definiendo el número de factores (es decir, las posiciones de las variables), el número de niveles (es decir, las opciones en cada posición) y el número de experimentos a realizar para obtener una matriz de salida. El contenido de información de la matriz de salida (que consiste típicamente de 1s y 0s que representan opciones de residuos en cada posición) depende directamente del número de experimentos a realizar (típicamente, cuantos más, mejor).
3) Usar la matriz de salida para construir una alineación de proteínas que codifique los 1s y 0s de vuelta a elecciones específicas de residuos en cada posición.
4) Sintetizar los genes que codifican las proteínas representadas en el alineamiento de proteínas.
5) Probar las proteínas codificadas por los genes sintetizados en ensayo o ensayos relevantes.
6) Construir un modelo basado en los genes/proteínas analizados.
7) Seguir los pasos descritos en la presente para identificar posiciones de importancia y construir una o más bibliotecas posteriores con una idoneidad mejorada.

En un ejemplo ilustrativo, se investiga una proteína en la que deben determinarse los residuos de aminoácidos funcionalmente mejores en 20 posiciones (por ejemplo, cuando hay 2 aminoácidos posibles disponibles en cada posición). En este ejemplo, sería apropiado un diseño factorial IV de resolución. Un diseño IV de resolución se define como un diseño capaz de dilucidar los efectos de todas las variables individuales, sin que se superpongan efectos de dos factores. El diseño especificaría entonces un conjunto de 40 secuencias específicas de aminoácidos que cubren

la diversidad total de 2^{20} (~1 millón) de secuencias posibles. A continuación, estas secuencias se generan usando cualquier protocolo estándar de síntesis de genes y se determina la función y la idoneidad de estos clones.

Una alternativa a los enfoques anteriores es emplear algunas o todas las secuencias disponibles (por ejemplo, la base de datos GENBANK® y otras fuentes públicas) para proporcionar la biblioteca de variantes de proteínas. Este enfoque proporciona una indicación de las regiones del espacio de secuencias de interés.

C. MÉTODOS DE SECUENCIACIÓN

Históricamente, la secuenciación ha sido un paso limitante en el desarrollo de grandes conjuntos de entrenamiento y, en consecuencia, de modelos secuencia-actividad cada vez más robustos. El elevado coste y el largo tiempo requeridos para secuenciar las variantes limitaban el número de observaciones a unas pocas decenas de variantes. Las herramientas de secuenciación de próxima generación han reducido enormemente el coste y han aumentado la velocidad y el volumen de secuenciación, permitiendo incluir variantes tanto de baja como de alta actividad en un conjunto de entrenamiento.

Las herramientas de secuenciación de próxima generación pueden secuenciar de manera económica grandes cantidades de pares de bases (por ejemplo, por lo menos aproximadamente 1.000.000.000 de pares de bases) en una sola serie. Esta capacidad puede utilizarse al secuenciar proteínas variantes, que típicamente tienen sólo unos pocos pares de kilobases de longitud, en una única serie. A menudo, las herramientas de secuenciación de próxima generación están optimizadas para secuenciar genomas únicos de gran tamaño (por ejemplo, el genoma humano) en lugar de muchas secuencias más pequeñas en una sola serie. Para aprovechar el potencial de las herramientas de secuenciación de próxima generación para secuenciar muchas observaciones en paralelo, debe identificarse de manera única el origen de cada una de las observaciones que se secuencian en una única serie. En algunas realizaciones, se usan secuencias con códigos de barras en todos y cada uno de los fragmentos alimentados a un secuenciador de próxima generación para una única serie. En un ejemplo, los códigos de barras identifican de manera única un pocillo concreto de una placa particular (por ejemplo, placas de 96 pocillos). En algunas de estas realizaciones, cada pocillo de cada placa contiene una única variante única. Mediante la codificación con códigos de barras de cada variante, o más específicamente de cada fragmento de cada variante, pueden secuenciarse e identificarse en una única serie las secuencias genéticas de múltiples variantes diferentes. En el proceso, todas las lecturas de fragmentos que tienen el mismo código de barras se identifican y procesan juntas por el algoritmo que identifica las secuencias de longitud de las variantes.

En algunas realizaciones, se extrae el ADN de las células de una variante en un pocillo dado y luego se fragmenta. A continuación, los fragmentos se codifican con códigos de barras para identificar por lo menos el pocillo y, a veces, el pocillo y la placa asociados a esa variante. Los fragmentos resultantes se seleccionan por tamaño para producir secuencias de longitud adecuada para el secuenciador de próxima generación. En un ejemplo ilustrativo, las longitudes de lectura son de aproximadamente 200 pares de bases. En algunas realizaciones, el código de barras de la placa no se aplica hasta después de que se hayan agrupado los fragmentos de ADN de los distintos pocillos de una placa. A continuación, se aplica un código de barras al ADN agrupado para identificar la placa. En algunas realizaciones, cada fragmento, independientemente del pocillo del que proceda, tendrá el mismo código de barras de la placa. Sin embargo, en algunas realizaciones alternativas, los fragmentos tienen códigos de barras diferentes. Además, pueden aplicarse los códigos de barras del pocillo y de la placa para identificar el ADN extraído de un pocillo dado.

En una o más realizaciones, los datos de la secuencia pueden obtenerse usando métodos de secuenciación masiva que incluyen, por ejemplo, la secuenciación de Sanger o la secuenciación de Maxam-Gilbert, que se consideran los métodos de secuenciación de primera generación. La secuenciación de Sanger, que implica el uso de terminadores de cadena dideoxi marcados, es bien conocida en la técnica; consultar, por ejemplo, Sanger et al., Proceedings of the National Academy of Sciences of the United States of America 74, 5463-5467 (1997). La secuenciación de Maxam-Gilbert, que implica la realización de múltiples reacciones de degradación química parcial en fracciones de la muestra de ácido nucleico seguido de la detección y el análisis de los fragmentos para inferir la secuencia, también es bien conocida en la técnica; consultar, por ejemplo, Maxam et al., Proceedings of the National Academy of Sciences of the United States of America 74, 560-564 (1977). Otro método de secuenciación masiva es la secuenciación por hibridación, en la que la secuencia de una muestra se deduce basándose en sus propiedades de hibridación con una pluralidad de secuencias, por ejemplo, en una micromatriz o chip de genes; consultar, por ejemplo, Drmanac, et al., Nature Biotechnology 16, 54-58 (1998).

En una o más realizaciones, los datos de secuencia se obtienen usando métodos de secuenciación de próxima generación. La secuenciación de próxima generación también se conoce como "secuenciación de alto rendimiento". Las técnicas paralelizan el proceso de secuenciación, produciendo miles o millones de secuencias a la vez. Algunos ejemplos de métodos de secuenciación de próxima generación adecuados incluyen, entre otros, la secuenciación en tiempo real de una sola molécula (por ejemplo, Pacific Biosciences, Menlo Park, California), la secuenciación por semiconductores iónicos (por ejemplo, Ion Torrent, South San Francisco, California), la pirosecuenciación (por ejemplo, 454, Branford, Connecticut), secuenciación por ligadura (por ejemplo, secuenciación

SOLid de Life Technologies, Carlsbad, California), secuenciación por síntesis y terminador reversible (por ejemplo, Illumina, San Diego, California), tecnologías de imagenología de ácidos nucleicos como la microscopía electrónica de transmisión, y similares.

En general, los métodos de secuenciación de próxima generación usan típicamente un paso de clonación in vitro para amplificar moléculas de ADN individuales. La PCR en emulsión (emPCR) aísla moléculas de ADN individuales junto con perlas recubiertas de cebadores en gotitas acuosas dentro de una fase oleosa. La PCR produce copias de la molécula de ADN, que se unen a los cebadores en la perla, seguidas de inmovilización para su posterior secuenciación. La emPCR se usa en los métodos de Marguilis et al. (comercializado por 454 Life Sciences, Branford, CT), Shendure y Porreca et al. (también conocido como "secuenciación polónica") y secuenciación SOLiD, (Applied Biosystems Inc., Foster City, CA). Consultar M. Margulies, et al. (2005) "Genome sequencing in microfabricated high-density picoliter reactors" Nature 437: 376-380; J. Shendure, et al. (2005) "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome" Science 309 (5741): 1728-1732. La amplificación clonal in vitro también puede llevarse a cabo mediante "PCR puente", en la que los fragmentos se amplifican sobre cebadores adheridos a una superficie sólida. Braslaysky et al. desarrollaron un método de molécula única (comercializado por Helicos Biosciences Corp., Cambridge, Mass.) que omite este paso de amplificación, fijando directamente las moléculas de ADN a una superficie. I. Braslaysky, et al. (2003) "Sequence information can be obtained from single DNA molecules" Proceedings of the National Academy of Sciences of the United States of America 100: 3960-3964.

Las moléculas de ADN unidas físicamente a una superficie pueden secuenciarse en paralelo. En la "secuenciación por síntesis", se construye una cadena complementaria sobre la base de la secuencia de una cadena plantilla usando una ADN polimerasa como en la secuenciación electroforética con terminación por colorante. Los métodos de terminadores reversibles (comercializado por Illumina, Inc., San Diego, CA, y Helicos Biosciences Corp., Cambridge, MA) usan versiones reversibles de colorantes terminadores, añadiendo un nucleótido cada vez, y detectan la fluorescencia en cada posición en tiempo real, mediante la eliminación repetida del grupo bloqueador para permitir la polimerización de otro nucleótido. La "pirosecuenciación" también usa la polimerización del ADN, añadiendo un nucleótido cada vez y detectando y cuantificando el número de nucleótidos añadidos a una posición dada a través de la luz emitida por la liberación de pirofosfatos unidos (comercializado por 454 Life Sciences, Branford, Conn.). Consultar M. Ronaghi, et al. (1996). "Real-time DNA sequencing using detection of pyrophosphate release" Analytical Biochemistry 242: 84-89.

A continuación se describen con más detalle ejemplos específicos de métodos de secuenciación de próxima generación. Una o más implementaciones de la presente invención pueden usar uno o más de los siguientes métodos de secuenciación sin desviarse de los principios de la invención.

La secuenciación en tiempo real de molécula única (también conocida como SMRT) es una tecnología paralelizada de secuenciación de ADN de molécula única por síntesis desarrollada por Pacific Biosciences. La secuenciación de molécula única en tiempo real utiliza la guía de ondas de modo cero (ZMW). Una única enzima ADN polimerasa se fija en la parte inferior de una ZMW con una única molécula de ADN como plantilla. La ZMW es una estructura que crea un volumen de observación iluminado lo suficientemente pequeño como para observar un solo nucleótido de ADN (también conocido como base) incorporado por la ADN polimerasa. Cada una de las cuatro bases de ADN está unida a uno de los cuatro colorantes fluorescentes diferentes. Cuando la ADN polimerasa incorpora un nucleótido, la etiqueta fluorescente se escinde y se difunde fuera del área de observación de la ZMW, donde su fluorescencia deja de ser observable. Un detector detecta la señal fluorescente de la incorporación del nucleótido, y la llamada de base se realiza según la fluorescencia correspondiente del colorante.

Otra tecnología de secuenciación de molécula única aplicable es la tecnología secuenciación de molécula única verdadera de Helicos (tSMS) (por ejemplo, como se describe en Harris T. D. et al., Science 320:106-109 [2008]). En la técnica tSMS, una muestra de ADN se escinde en cadenas de aproximadamente 100 a 200 nucleótidos, y se añade una secuencia poliA al extremo 3′ de cada cadena de ADN. Cada cadena se marca mediante la adición de un nucleótido de adenosina marcado con fluorescencia. A continuación, las cadenas de ADN se hibridan en una célula de flujo, que contiene millones de sitios de captura de oligo-T inmovilizados en la superficie de la célula de flujo. En ciertas realizaciones, las plantillas pueden tener una densidad de aproximadamente 100 millones de plantillas/cm^2. A continuación, la celda de flujo se carga en un instrumento, por ejemplo, el secuenciador HeliScope™, y un láser ilumina la superficie de la celda de flujo, revelando la posición de cada plantilla. Una cámara CCD puede mapear la posición de las plantillas en la superficie de la celda de flujo. A continuación, se escinde y se lava el marcador fluorescente de la plantilla. La reacción de secuenciación comienza introduciendo una ADN polimerasa y un nucleótido marcado con fluorescencia. El ácido nucleico oligo-T sirve como cebador. La polimerasa incorpora los nucleótidos marcados al cebador de forma dirigida a la plantilla. La polimerasa y los nucleótidos no incorporados se eliminan. Las plantillas que han incorporado directamente el nucleótido marcado con fluorescencia se distinguen mediante imagenología de la superficie de la célula de flujo. Después de la imagenología, un paso de escisión elimina el marcador fluorescente y el proceso se repite con otros nucleótidos marcados con fluorescencia hasta que se alcanza la longitud de lectura deseada. La información de la secuencia se recoge con cada paso de adición de nucleótidos. La secuenciación del genoma completo mediante tecnologías de secuenciación de molécula única excluye o normalmente obvia la amplificación basada en PCR en la preparación de las bibliotecas de secuenciación, y los métodos permiten la

medición directa de la muestra, en lugar de la medición de copias de esa muestra.

La secuenciación por iones semiconductores es un método de secuenciación del ADN basado en la detección de iones de hidrógeno que se liberan durante la polimerización del ADN. Se trata de un método de "secuenciación por síntesis", durante el cual se construye una cadena complementaria sobre la base de la secuencia de una cadena plantilla. Un micropocillo que contiene una cadena plantilla de ADN que se va a secuenciar se inunda con una única especie de desoxirribonucleótido trifosfato (dNTP). Si el dNTP introducido es complementario al nucleótido principal de la plantilla, se incorpora a la cadena complementaria en crecimiento. Esto provoca la liberación de un ion hidrógeno que activa un sensor de iones ISFET, lo que indica que se ha producido una reacción. Si hay repeticiones de homopolímeros en la secuencia plantilla, se incorporarán múltiples moléculas de dNTP en un solo ciclo. Esto lleva a un número correspondiente de hidrógenos liberados y a una señal electrónica proporcionalmente mayor. Esta tecnología difiere de otras tecnologías de secuenciación en que no se usan nucleótidos modificados u ópticos. La secuenciación por ion semiconductor también puede denominarse secuenciación por torrente iónico, secuenciación mediada por pH, secuenciación por silicio o secuenciación por semiconductor.

En la pirosecuenciación, el ion pirofosfato liberado por la reacción de polimerización se hace reaccionar con adenosina 5′ fosfosulfato por la ATP sulfurilasa para producir ATP; el ATP impulsa entonces la conversión de luciferina en oxiluciferina más luz por la luciferasa. Como la fluorescencia es transitoria, en este método no es necesario ningún paso separado para eliminar la fluorescencia. Se añade un tipo de desoxirribonucleótido trifosfato (dNTP) cada vez, y la información de la secuencia se discierne de acuerdo con el dNTP que genera una señal significativa en un lugar de la reacción. El instrumento GS FLX de Roche, disponible en el mercado, adquiere la secuencia usando este método. Esta técnica y sus aplicaciones se tratan en detalle, por ejemplo, en Ronaghi et al., Analytical Biochemistry 242, 84-89 (1996) y Margulies et al., Nature 437, 376-380 (2005) (corrigendum en Nature 441, 120 (2006)). Una tecnología de pirosecuenciación disponible comercialmente es la secuenciación 454 (Roche) (por ejemplo, como se describe en Margulies, M. et al. Nature 437:376-380 [2005]).

En la secuenciación por ligación, se usa una enzima ligasa para unir un oligonucleótido parcialmente de cadena doble con un saliente al ácido nucleico que se está secuenciando, que tiene un saliente; para que se produzca la ligación los salientes deben ser complementarios. Las bases del saliente del oligonucleótido parcialmente de cadena doble pueden identificarse de acuerdo con un fluoróforo conjugado con el oligonucleótido parcialmente de cadena doble y/o con un oligonucleótido secundario que se hibrida con otra parte del oligonucleótido parcialmente de cadena doble. Tras la adquisición de los datos de fluorescencia, el complejo ligado se escinde en sentido ascendente del sitio de ligación, por ejemplo, mediante una enzima de restricción de tipo IIs, por ejemplo, Bbvl, que corta en un sitio a una distancia fija de su sitio de reconocimiento (que se incluyó en el oligonucleótido parcialmente de cadena doble). Esta reacción de escisión expone un nuevo saliente justo en sentido ascendente del anterior, y el proceso se repite. Esta técnica y sus aplicaciones se analizan en detalle, por ejemplo, en Brenner et al., Nature Biotechnology 18, 630-634 (2000). En algunas realizaciones, la secuenciación por ligadura se adapta a los métodos de la invención obteniendo un producto de amplificación de círculo rodante de una molécula de ácido nucleico circular, y usando el producto de amplificación de círculo rodante como plantilla para la secuenciación por ligadura.

Un ejemplo disponible comercialmente de tecnología de secuenciación por ligación es la tecnología SOLiD™ (Applied Biosystems). En la secuenciación por ligación SOLiD™, el ADN genómico se corta en fragmentos y los adaptadores se unen a los extremos 5′ y 3′ de los fragmentos para generar una biblioteca de fragmentos. Alternativamente, pueden introducirse adaptadores internos ligando adaptadores a los extremos 5′ y 3′ de los fragmentos, circularizando los fragmentos, digiriendo el fragmento circularizado para generar un adaptador interno y uniendo adaptadores a los extremos 5′ y 3′ de los fragmentos resultantes para generar una biblioteca apareada. A continuación, se preparan poblaciones clonales de perlas en microrreactores que contienen perlas, cebadores, plantillas y componentes de PCR. Tras la PCR, las plantillas se desnaturalizan y las perlas se enriquecen para separar las perlas con plantillas extendidas. Las plantillas en las perlas seleccionadas se someten a una modificación 3′ que permite su unión a un portaobjetos de vidrio. La secuencia puede determinarse por hibridación secuencial y ligación de oligonucleótidos parcialmente aleatorios con una base (o par de bases) central determinada que se identifica mediante un fluoróforo específico. Después de registrar un color, el oligonucleótido ligado se escinde y se retira y, a continuación, se repite el proceso.

En la secuenciación con terminador reversible, se incorpora a una reacción de extensión de única base un análogo de nucleótido marcado con un colorante fluorescente que es un terminador de cadena reversible debido a la presencia de un grupo bloqueador. La identidad de la base se determina en función del fluoróforo; en otras palabras, cada base se empareja con un fluoróforo diferente. Una vez adquiridos los datos de fluorescencia/secuencia, se eliminan químicamente el fluoróforo y el grupo de bloqueo, y se repite el ciclo para adquirir la siguiente base de información de secuencia. El instrumento GA de Illumina funciona con este método. Esta técnica y sus aplicaciones se analizan con detalle, por ejemplo, en Ruparel et al., Proceedings of the National Academy of Sciences of the United States of America 102, 5932-5937 (2005), y Harris et al., Science 320, 106-109 (2008).

Un ejemplo disponible comercialmente de método de secuenciación con terminador reversible es la secuenciación por síntesis y la secuenciación basada en terminador reversible de Illumina (por ejemplo, como se

describe en Bentley et al., Nature 6:53-59 [2009]). La tecnología de secuenciación de Illumina se basa en la fijación de ADN genómico fragmentado a una superficie plana ópticamente transparente a la que se unen anclajes de oligonucleótidos. El ADN plantilla se repara en sus extremos para generar extremos romos fosforilados en 5′, y se usa la actividad polimerasa del fragmento de Klenow para añadir una única base A al extremo 3′ de los fragmentos de ADN fosforilados romos. Esta adición prepara los fragmentos de ADN para la ligación a adaptadores de oligonucleótidos, que tienen un saliente de una sola base T en su extremo 3′ para aumentar la eficiencia de la ligación. Los oligonucleótidos adaptadores son complementarios a los anclajes de la celda de flujo. En condiciones de dilución limitativa, se añade el ADN plantilla de cadena sencilla modificado por el adaptador a la celda de flujo y se inmoviliza mediante hibridación con los anclajes. Los fragmentos de ADN unidos se extienden y amplifican por puente para crear una celda de flujo de secuenciación de densidad ultra alta con cientos de millones de grupos, cada uno de los cuales contiene ~1.000 copias de la misma plantilla. Las plantillas se secuencian usando una tecnología de secuenciación por síntesis de ADN en cuatro colores robusta que emplea terminadores reversibles con colorantes fluorescentes extraíbles. La detección por fluorescencia de alta sensibilidad se consigue mediante excitación láser y óptica de reflexión interna total. Las lecturas de secuencias cortas de aproximadamente 20-40 pb, por ejemplo 36 pb, se alinean con un genoma de referencia enmascarado de repeticiones y se identifica la correspondencia única de las lecturas de secuencias cortas con el genoma de referencia mediante un software de análisis de datos especialmente desarrollado. También pueden usarse genomas de referencia sin enmascaramiento de repeticiones. Tanto si se usan genomas de referencia con máscara de repetición como sin ella, sólo se cuentan las lecturas que se corresponden de manera única con el genoma de referencia. Una vez finalizada la primera lectura, las plantillas pueden regenerarse in situ para permitir una segunda lectura desde el extremo opuesto de los fragmentos. Por tanto, puede usarse la secuenciación de extremo único o de extremo emparejado de los fragmentos de ADN. Se realiza la secuenciación parcial de los fragmentos de ADN presentes en la muestra, y se cuentan las etiquetas de secuencia que comprenden lecturas de longitud predeterminada, por ejemplo, 36 pb, se mapean a un genoma de referencia conocido.

En la secuenciación de nanoporos, se hace pasar una molécula de ácido nucleico de cadena sencilla a través de un poro, por ejemplo, usando una fuerza electroforética motriz, y la secuencia se deduce analizando los datos obtenidos a medida que la molécula de ácido nucleico de cadena sencilla atraviesa el poro. Los datos pueden ser datos de corriente iónica, en los que cada base altera la corriente, por ejemplo, bloqueando parcialmente la corriente que pasa a través del poro en un grado diferente y distinguible.

En otra realización ilustrativa, pero no limitativa, los métodos descritos en la presente comprenden la obtención de información de secuencia mediante microscopía electrónica de transmisión (TEM). El método comprende utilizar imagenología de microscopía electrónica de transmisión con resolución de átomo único de ADN de alto peso molecular (150 kb o más) marcado selectivamente con marcadores de átomo pesado y la disposición de estas moléculas en películas ultrafinas en matrices paralelas ultradensas (3 nm cadena a cadena) con espaciado consistente de base a base. El microscopio electrónico se usa para obtener imágenes de las moléculas en las películas para determinar la posición de los marcadores de átomos pesados y para extraer información de la secuencia de bases del ADN. El método se describe con más detalle en la publicación de patente PCT WO 2009/046445.

En otra realización ilustrativa, pero no limitativa, los métodos descritos en la presente comprenden la obtención de información de secuencias mediante secuenciación de tercera generación. En la secuenciación de tercera generación, se usa un portaobjetos con un recubrimiento de aluminio con muchos orificios pequeños (~ 50 nm) como guía de ondas de modo cero (consultar, por ejemplo, Levene et al., Science 299, 682-686 (2003)). La superficie de aluminio está protegida de la adhesión de la ADN polimerasa mediante química de polifosfonato, por ejemplo, química de polivinilfosfonato (consultar, por ejemplo, Korlach et al., Proceedings of the National Academy of Sciences of the United States of America 105, 1176-1181 (2008)). De este modo, las moléculas de ADN polimerasa se adhieren preferentemente a la sílice expuesta en los orificios del recubrimiento de aluminio. Esta configuración permite usar fenómenos de ondas evanescentes para reducir el fondo de fluorescencia, permitiendo el uso de mayores concentraciones de dNTPs marcados fluorescentemente. El fluoróforo está unido al fosfato terminal de los dNTPs, de tal manera que la fluorescencia se libera tras la incorporación del dNTP, pero el fluoróforo no permanece unido al nucleótido recién incorporado, lo que significa que el complejo está inmediatamente listo para otra ronda de incorporación. Mediante este método, puede detectarse la incorporación de dNTPs en complejos cebador-plantilla individuales presentes en los orificios del recubrimiento de aluminio. Ver, por ejemplo, Eid et al., Science 323, 133-138 (2009).

D. GENERACIÓN DE UN MODELO SECUENCIA-ACTIVIDAD

Como se ha indicado anteriormente, un modelo de secuencia-actividad usado con las realizaciones del presente documento relaciona la información de la secuencia de la proteína con la actividad de la proteína. La información de la secuencia de la proteína usada por el modelo puede adoptar muchas formas. En algunas realizaciones, es una secuencia completa de los residuos de aminoácidos de una proteína (por ejemplo, HGPVFSTGGA...). Sin embargo, en algunas realizaciones, no es necesaria la secuencia completa de aminoácidos. Por ejemplo, en algunas realizaciones, es suficiente proporcionar sólo aquellos residuos que van a ser variados en un esfuerzo de investigación particular. En algunas realizaciones que implican etapas posteriores de investigación, muchos residuos son fijos y sólo quedan por explorar regiones limitadas del espacio de secuencia. En algunas de tales

situaciones, es conveniente proporcionar modelos de secuencia-actividad que requieran, como entradas, sólo la identificación de aquellos residuos en las regiones de la proteína donde la exploración continúa. En algunas realizaciones adicionales, los modelos no requieren que se conozcan las identidades exactas de los residuos en las posiciones de los residuos. En algunas realizaciones, se identifican una o más propiedades físicas o químicas que caracterizan el aminoácido en una posición de residuo particular. En un ejemplo ilustrativo, el modelo requiere la especificación de las posiciones de los residuos por volumen, hidrofobicidad, acidez, etc. Además, en algunos modelos, se emplean combinaciones de dichas propiedades. De hecho, no se pretende que la presente invención se limite a ningún enfoque en particular, ya que los modelos encuentran uso en varias configuraciones de información de secuencia, información de actividad y/u otras propiedades físicas (por ejemplo, hidrofobicidad, etc.).

Por tanto, la forma del modelo secuencia-actividad puede variar ampliamente, siempre que proporcione un vehículo para aproximar correctamente la actividad relativa de las proteínas basándose en la información de la secuencia, según se desee. En algunas realizaciones, los modelos generalmente tratan la actividad como una variable dependiente y los valores de secuencia/residuo como variables independientes. Ejemplos de la forma matemática/lógica de los modelos incluyen expresiones matemáticas lineales y no lineales de varios órdenes, redes neuronales, árboles/gráficos de clasificación y regresión, enfoques de agrupación, partición recursiva, máquinas de vectores de soporte, y similares. En una realización, la forma del modelo es un modelo aditivo lineal en el que se suman los productos de los coeficientes y los valores residuales. En otra realización, la forma del modelo es un producto no lineal de varios términos de secuencia/residuo, incluyendo ciertos productos cruzados de residuo (que representan términos de interacción entre residuos). De hecho, no se pretende que las realizaciones divulgadas se limiten a ningún formato específico, ya que cualquier formato adecuado encuentra uso, como se ilustra en la presente.

En algunas realizaciones, los modelos se desarrollan a partir de un conjunto de entrenamiento de información de actividad frente a secuencia para proporcionar la relación matemática/lógica entre actividad y secuencia. Esta relación se valida típicamente antes de su uso para predecir la actividad de nuevas secuencias o el impacto de los residuos sobre la actividad de interés.

Hay varias técnicas para generar modelos y que encuentran uso en la presente invención. En algunas realizaciones, las técnicas implican la optimización de modelos o la minimización de errores de modelo. Ejemplos específicos incluyen mínimos cuadrados parciales, regresión de conjunto, bosque aleatorio, varias otras técnicas de regresión, así como técnicas de redes neuronales, partición recursiva, técnicas de máquinas de vectores de soporte, CART (árboles de clasificación y regresión), y/o similares. Generalmente, la técnica debe producir un modelo que pueda distinguir los residuos que tienen un impacto significativo sobre la actividad de los que no lo tienen. En algunas realizaciones, los modelos también clasifican residuos individuales o posiciones de residuos sobre la base de su impacto en la actividad. No se pretende que la presente invención se limite a ninguna técnica específica para generar modelos, ya que en la presente invención encuentra uso cualquier método adecuado conocido en la técnica.

En algunas realizaciones, los modelos se generan mediante una técnica de regresión que identifica la covariación de variables independientes y dependientes en un conjunto de entrenamiento. Se conocen y se usan ampliamente varias técnicas de regresión. Algunos ejemplos son la regresión lineal múltiple (MLR), la regresión de componentes principales (PCR) y la regresión de mínimos cuadrados parciales (PLS). En algunas realizaciones, los modelos se generan usando técnicas que implican múltiples componentes, incluyendo pero no limitados a la regresión de conjunto y el bosque aleatorio. En la presente invención se usan estos y otros métodos adecuados. No se pretende que la presente invención se limite a ninguna técnica en particular.

La MLR es la más básica de estas técnicas. Se usa simplemente para resolver un conjunto de ecuaciones de coeficientes para los miembros de un conjunto de entrenamiento. Cada ecuación se refiere a la actividad de un miembro del conjunto de entrenamiento (es decir, variables dependientes) con la presencia o ausencia de un residuo concreto en una posición determinada (es decir, variables independientes). Dependiendo del número de opciones de residuos en el conjunto de entrenamiento, el número de estas ecuaciones puede ser bastante grande.

Al igual que la MLR, la PLS y la PCR generan modelos a partir de ecuaciones que relacionan la actividad de la secuencia con los valores de los residuos. Sin embargo, estas técnicas lo hacen de manera diferente. Primero realizan una transformación de coordenadas para reducir el número de variables independientes. A continuación, realizan la regresión sobre las variables transformadas. En la MLR, existe un número potencialmente muy elevado de variables independientes: dos o más por cada posición de residuo que varía dentro del conjunto de entrenamiento. Dado que las proteínas y los péptidos de interés son a menudo bastante grandes y que el conjunto de entrenamiento puede proporcionar muchas secuencias diferentes, el número de variables independientes puede llegar a ser rápidamente muy grande. Al reducir el número de variables para centrarse en las que proporcionan la mayor variación en el conjunto de datos, PLS y PCR requieren generalmente menos muestras y simplifican los pasos necesarios para generar modelos.

La PCR es similar a la regresión PLS en que la regresión real se realiza sobre un número relativamente pequeño de variables latentes obtenidas por transformación de coordenadas de las variables independientes brutas (es decir, los valores de residuos). La diferencia entre PLS y PCR es que las variables latentes en PCR se construyen

maximizando la covariación entre las variables independientes (es decir, los valores de residuos). En la regresión PLS, las variables latentes se construyen maximizando la covariación entre las variables independientes y las variables dependientes (es decir, los valores de actividad). La regresión por mínimos cuadrados parciales se describe en Hand, D. J., et al. (2001) Principles of Data Mining (Adaptive Computation and Machine Learning), Boston, Mass., MIT Press, y en Geladi, et al. (1986) "Partial Least-Squares Regression: a Tutorial", Analytica Chimica Acta, 198:1-17.

En la PCR y la PLS, el resultado directo del análisis de regresión es una expresión para la actividad que es una función de las variables latentes ponderadas. Esta expresión puede transformarse en una expresión de la actividad en función de las variables independientes originales realizando una transformación de coordenadas que convierta las variables latentes de nuevo en las variables independientes originales.

En esencia, tanto la PCR como la PLS reducen primero la dimensionalidad de la información contenida en el conjunto de entrenamiento y luego realizan un análisis de regresión sobre un conjunto de datos transformado, que se ha transformado para producir nuevas variables independientes, pero conserva los valores originales de la variable dependiente. Las versiones transformadas de los conjuntos de datos pueden dar lugar a un número relativamente reducido de expresiones para realizar el análisis de regresión. En los protocolos en los que no se ha realizado ninguna reducción de dimensión, debe considerarse cada residuo separado para el que puede haber una variación. Esto puede suponer un conjunto muy grande de coeficientes (por ejemplo, 2^N coeficientes para interacciones bidireccionales, donde N es el número de posiciones de residuos que pueden variar en el conjunto de entrenamiento). En un análisis típico de componentes principales, sólo se emplean 3, 4, 5, 6 componentes principales.

La capacidad de las técnicas de aprendizaje automático para ajustarse a los datos de entrenamiento se denomina a menudo "ajuste del modelo" y, en técnicas de regresión como MLR, PCR y PLS, el ajuste del modelo se mide típicamente por la diferencia cuadrática entre los valores medidos y los predichos. Para un conjunto de entrenamiento determinado, el ajuste óptimo del modelo se conseguirá usando MLR, mientras que PCR y PLS tienen a menudo un peor ajuste del modelo (mayor error cuadrático entre las mediciones y las predicciones). Sin embargo, la principal ventaja de usar técnicas de regresión de variables latentes como PCR y PLS reside en la capacidad predictiva de dichos modelos. La obtención de un ajuste del modelo con un error cuadrático sumatorio muy pequeño no garantiza en modo alguno que el modelo sea capaz de predecir con exactitud nuevas muestras no observadas en el conjunto de entrenamiento; de hecho, a menudo se produce lo contrario, sobre todo cuando hay muchas variables y sólo unas pocas observaciones (es decir, muestras). Por tanto, las técnicas de regresión de variables latentes (por ejemplo, PCR, PLS), aunque a menudo tienen peores ajustes del modelo en los datos de entrenamiento, habitualmente son más robustas y son capaces de predecir con mayor precisión nuevas muestras fuera del conjunto de entrenamiento.

Otra clase de herramientas que pueden usarse para generar modelos de acuerdo con la presente divulgación son las máquinas de vectores de soporte (SVM). Estas herramientas matemáticas toman como entradas conjuntos de secuencias de entrenamiento que se han clasificado en dos o más grupos en función de la actividad. Las máquinas de vectores de soporte funcionan ponderando de manera diferente los distintos miembros de un conjunto de entrenamiento en función de lo cerca que estén de una interfaz hiperplana que separa los miembros "activos" e "inactivos" del conjunto de entrenamiento. Esta técnica requiere que el científico decida primero qué miembros del conjunto de entrenamiento colocar en el grupo "activo" y qué miembros del conjunto de entrenamiento colocar en el grupo "inactivo". En algunas realizaciones, esto se consigue eligiendo un valor numérico apropiado para el nivel de actividad que sirve como límite entre los miembros "activos" e "inactivos" del conjunto de entrenamiento. A partir de esta clasificación, la máquina de vectores de soporte genera un vector, W, que puede proporcionar valores de coeficiente para las variables independientes individuales que definen las secuencias de los miembros activos e inactivos del grupo en el conjunto de entrenamiento. Estos coeficientes pueden usarse para "clasificar" los residuos individuales, como se describe en otra parte de la presente. La técnica se usa para identificar un hiperplano que maximiza la distancia entre los miembros del conjunto de entrenamiento más cercanos en lados opuestos de ese plano. En otra realización, se lleva a cabo el modelado de regresión de vectores de soporte. En este caso, la variable dependiente es un vector de valores continuos de actividad. El modelo de regresión de vectores de soporte genera un vector de coeficientes, W, que puede usarse para clasificar residuos individuales.

Las SVM se han usado para analizar grandes conjuntos de datos en muchos estudios y han encontrado un amplio uso en los micromatrices de ADN. Entre sus puntos fuertes potenciales se encuentra la capacidad de discriminar con precisión (mediante ponderación) los factores que separan unas muestras de otras. En la medida en que una SVM puede desentrañar con precisión qué residuos contribuyen a la función, puede ser una herramienta particularmente útil para clasificar los residuos. Las SVM se describen en S. Gunn (1998) "Support Vector Machines for Classification and Regressions," Technical Report, Faculty of Engineering and Applied Science, Department of Electronics and Computer Science, University of Southampton.

En algunas realizaciones de la invención, otra clase de herramientas que pueden usarse para generar modelos son la clasificación y la regresión basadas en un conjunto de árboles de clasificación que usan entradas aleatorias, un ejemplo de las cuales es el bosque aleatorio. Ver Breiman (2001). "Random Forests", Machine Learning 45 (1): 5-32. Los bosques aleatorios son una combinación de predictores de árboles de tal manera que cada árbol

depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles del bosque. Un bosque aleatorio es un conjunto de aprendizaje formado por un grupo de aprendices de árboles de decisión no podados con una selección aleatoria de características en cada división del árbol de decisión. El error de generalización de los bosques converge a un límite a medida que aumenta el número de árboles del bosque.

Los bosques aleatorios pueden construirse de la siguiente manera:

1) Si el número de casos del conjunto de entrenamiento es N, muestrear N casos aleatoriamente, pero con reemplazo, de los datos originales. Esta muestra será el conjunto de entrenamiento para el crecimiento del árbol.
2) Si hay M variables independientes de entrada, se especifica un número m<<M tal que en cada nodo del árbol se seleccionan aleatoriamente m variables de entre las M y se usa la mejor división de estas m para dividir el nodo. El valor de m se mantiene constante durante el crecimiento del bosque.
3) En algunas implementaciones, cada árbol crece lo máximo posible. No hay poda.
4) A continuación se genera un gran número de árboles, k = 1,..., K (normalmente K>=100).
5) Después de haber generado un gran número de árboles, todos ellos votan para la clasificación de las variables de interés. Por ejemplo, cada uno de ellos puede contribuir a la predicción final de actividad o a la contribución de mutaciones particulares.
6) A continuación, el bosque aleatorio clasifica x (por ejemplo, una secuencia de mutaciones u otra variable independiente) tomando la clase más votada de entre todos los árboles predictores del bosque.

La tasa de error del bosque depende de la correlación entre dos árboles cualesquiera del bosque. Aumentar la correlación incrementa la tasa de error del bosque. La tasa de error del bosque depende de la fuerza de cada árbol del bosque. Un árbol con una tasa de error baja es un clasificador fuerte. Aumentar la fuerza de los árboles individuales disminuye la tasa de error del bosque. Reducir m reduce tanto la correlación como la fuerza. Si se aumenta, aumentan ambas. En algún punto intermedio se encuentra el intervalo "óptimo" de m, que habitualmente es bastante amplio.

Las técnicas de bosque aleatorio pueden usarse tanto para variables categóricas como para variables continuas en modelos de regresión. En algunas realizaciones de la invención, los modelos de bosque aleatorio tienen un potencia predictiva comparable a los modelos SVM y de red neuronal, pero tienden a tener una mayor eficiencia computacional porque, entre otras razones, la validación cruzada está integrada en el proceso de modelado y no es necesario un proceso separado para la validación cruzada.

i) Modelos lineales

Aunque la presente divulgación se dirige a modelos no lineales, éstos pueden entenderse más fácilmente en el contexto de modelos lineales de secuencia frente a actividad. Además, en algunas realizaciones, un modelo lineal se utiliza como modelo "base" en un proceso gradual para generar un modelo no lineal. En general, un modelo de regresión lineal de actividad frente a secuencia tiene la siguiente forma:

$$y = c_0 + \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} X_{ij} \qquad (1)$$

En esta expresión lineal, y es la respuesta predicha, mientras que c_{ij} y x_{ij} son el coeficiente de regresión y el valor de bit o variable ficticia usados para representar la elección de residuo, respectivamente en la posición i de la secuencia. Hay N posiciones de residuos en las secuencias de la biblioteca de variantes de proteínas y cada una de ellas puede estar ocupada por uno o más residuos. En cualquier posición, puede haber de j=1 a M tipos de residuos distintos. Este modelo asume una relación lineal (aditiva) entre los residuos en cada posición. A continuación se presenta una versión ampliada de la ecuación 1:

$$y=c_0 + c_{11}x_{11} + c_{12}x_{12} + \ldots c_{1M}x_{1M} + c_{21}x_{21} + c_{22}x_{22} + \ldots c_{2M}x_{2M} + \ldots + c_{NM}x_{NM}$$

Como se ha indicado, los datos en forma de información de actividad y secuencia se derivan de la biblioteca inicial de variantes de proteínas y se usan para determinar los coeficientes de regresión del modelo. Las variables ficticias se identifican primero a partir de un alineamiento de las secuencias de variantes de proteínas. Se identifican posiciones de residuos de aminoácidos entre las secuencias de variantes de proteínas en las que los residuos de aminoácidos en esas posiciones difieren entre secuencias. La información sobre residuos de aminoácidos en algunas o todas estas posiciones de residuos variables puede incorporarse al modelo secuencia-actividad.

La Tabla I contiene información sobre la secuencia en forma de posiciones de residuos variables y tipos de residuos para 10 proteínas variantes ilustrativas, junto con los valores de actividad correspondientes a cada proteína variante. Estos son miembros representativos de un conjunto mayor que se requiere para generar suficientes ecuaciones para resolver todos los coeficientes. Así, por ejemplo, para las secuencias variantes de proteínas

ilustrativas de la Tabla I, las posiciones 10, 166, 175 y 340 son posiciones de residuos variables y todas las demás posiciones, es decir, las que no se indican en la Tabla, contienen residuos que son idénticos entre las Variantes 1-10.

En este ejemplo, las 10 variantes pueden incluir o no la secuencia de estructura principal de tipo salvaje. En algunas realizaciones, un modelo desarrollado para que tenga en cuenta los datos de todas las variantes que incluyen la secuencia troncal de tipo salvaje puede introducir un problema de multicolinealidad perfecta, o una trampa de variable ficticia. Este problema puede abordarse mediante varias técnicas. Algunas realizaciones pueden excluir los datos de la estructura principal de tipo salvaje del desarrollo del modelo. Algunas realizaciones pueden descartar los coeficientes que representan la estructura principal de tipo salvaje. Algunas realizaciones pueden usar técnicas como la regresión PLS para abordar la multicolinealidad.

Tabla I: Secuencia ilustrativa y datos de actividad

Posición del residuo variable	10	166	175	340	y (actividad)
Variante 1	Ala	Ser	Gly	Phe	y_1
Variante 2	Asp	Phe	Val	Ala	y_2
Variante 3	Lys	Leu	Gly	Ala	y_3
Variante 4	Asp	Isla	Val	Phe	y_4
Variante 5	Ala	Isla	Val	Ala	y_5
Variante 6	Asp	Ser	Gly	Phe	y_6
Variante 7	Lys	Phe	Gly	Phe	y_7
Variante 8	Ala	Phe	Val	Ala	y_8
Variante 9	Lys	Ser	Gly	Phe	y_9
Variante 10	Asp	Leu	Val	Ala	y_{10}

Por tanto, basándose en la ecuación 1, puede derivarse un modelo de regresión a partir de la biblioteca sistemáticamente variada de la Tabla I, es decir:

$$y = c_0 + c_{10\,Ala} X_{10Ala} + c_{10Asp} X_{10Asp} + c_{10\,Lys} X_{10Lys} + c_{166Ser} X_{166Ser} + c_{166\,Phe} X_{166Phe} + c_{166Leu} X_{166Leu} + c_{166Ile} X_{166Ile} + c_{175Gly} X_{175Gly} + c_{175\,Val} X_{175Val} + c_{340\,Phe} X_{340Phe} + c_{340\,Ala} X_{340Ala} \quad (Ec.\ 2)$$

Los valores de bits (variables ficticias x) pueden representarse como 1 o 0, reflejando la presencia o ausencia del residuo de aminoácido designado o, alternativamente, 1 o -1, o alguna otra representación sustitutiva. Por ejemplo, usar la designación 1 o 0, x_{10Ala} sería "1" para la Variante 1 y "0" para la Variante 2. Usar la designación 1 o -1, x_{10Ala} sería "1" para la Variante 1 y "-1" para la Variante 2. De este modo, los coeficientes de regresión pueden derivarse de ecuaciones de regresión basadas en la información de actividad de secuencia para todas las variantes de la biblioteca. A continuación figuran ejemplos de tales ecuaciones para las variantes 1-10 (usando la designación 1 o 0 para x):

$$y_1 = c_0 + c_{10\,Ala}(1) + c_{10Asp}(0) + c_{10\,Lys}(0) + c_{166Ser}(1) + c_{166\,Phe}(0) + c_{166Leu}(0) + c_{166Ile}(0) + c_{175Gly}(1) + c_{175\,Val}(0) + c_{340\,Phe}(1) + c_{340\,Ala}(0)$$

$$y_2 = c_0 + c_{10\,Ala}(0) + c_{10Asp}(1) + c_{10\,Lys}(0) + c_{166Ser}(0) + c_{166\,Phe}(1) + c_{166Leu}(0) + c_{166Ile}(0) + c_{175Gly}(0) + c_{175\,Val}(1) + c_{340\,Phe}(0) + c_{340\,Ala}(1)$$

$$y_3 = c_0 + c_{10\,Ala}(0) + c_{10Asp}(0) + c_{10\,Lys}(1) + c_{166Ser}(0) + c_{166\,Phe}(0) + c_{166Leu}(1) + c_{166Ile}(0) + c_{175Gly}(1) + c_{175\,Val}(0) + c_{340\,Phe}(0) + c_{340\,Ala}(1)$$

$$y_4 = c_0 + c_{10\,Ala}(0) + c_{10Asp}(1) + c_{10\,Lys}(0) + c_{166Ser}(0) + c_{166\,Phe}(0) + c_{166Leu}(0) + c_{166Ile}(1) + c_{175Gly}(0) + c_{175\,Val}(1) + c_{340\,Phe}(1) + c_{340\,Ala}(0)$$

$$y_5 = c_0 + c_{10\ Ala}(1) + c_{10Asp}(0) + c_{10\ Lys}(0) + c_{166Ser}(0) + c_{166\ Phe}(0) + c_{166Leu}(0) + c_{166Ile}(1) + c_{175Gly}(0) + c_{175\ Val}(1) + c_{340\ Phe}(0) + c_{340\ Ala}(1)$$

$$y_6 = c_0 + c_{10\ Ala}(0) + c_{10Asp}(1) + c_{10\ Lys}(0) + c_{166Ser}(1) + c_{166\ Phe}(0) + c_{166Leu}(0) + c_{166Ile}(0) + c_{175Gly}(1) + c_{175\ Val}(0) + c_{340\ Phe}(1) + c_{340\ Ala}(0)$$

$$y_7 = c_0 + c_{10\ Ala}(0) + c_{10Asp}(0) + c_{10\ Lys}(1) + c_{166Ser}(0) + c_{166\ Phe}(1) + c_{166Leu}(0) + c_{166Ile}(0) + c_{175Gly}(1) + c_{175\ Val}(0) + c_{340\ Phe}(1) + c_{340\ Ala}(0)$$

$$y_8 = c_0 + c_{10\ Ala}(1) + c_{10Asp}(0) + c_{10\ Lys}(0) + c_{166Ser}(0) + c_{166\ Phe}(1) + c_{166Leu}(0) + c_{166Ile}(0) + c_{175Gly}(0) + c_{175\ Val}(1) + c_{340\ Phe}(0) + c_{340\ Ala}(1)$$

$$y_9 = c_0 + c_{10\ Ala}(0) + c_{10Asp}(0) + c_{10\ Lys}(1) + c_{166Ser}(1) + c_{166\ Phe}(0) + c_{166Leu}(0) + c_{166Ile}(0) + c_{175Gly}(1) + c_{175\ Val}(0) + c_{340\ Phe}(1) + c_{340\ Ala}(0)$$

$$y_{10} = c_0 + c_{10\ Ala}(0) + c_{10Asp}(1) + c_{10\ Lys}(0) + c_{166Ser}(0) + c_{166\ Phe}(0) + c_{166Leu}(1) + c_{166Ile}(0) + c_{175Gly}(0) + c_{175\ Val}(1) + c_{340\ Phe}(0) + c_{340\ Ala}(1)$$

El conjunto completo de ecuaciones puede resolverse fácilmente usando cualquier técnica de regresión adecuada (por ejemplo, PCR, PLS o MLR) para determinar el valor de los coeficientes de regresión correspondientes a cada residuo y posición de interés. En este ejemplo, la magnitud relativa del coeficiente de regresión se correlaciona con la magnitud relativa de la contribución de ese residuo particular en la posición particular a la actividad. Los coeficientes de regresión pueden entonces clasificarse o categorizarse de otro modo para determinar qué residuos tienen más probabilidades de contribuir favorablemente a la actividad deseada. La Tabla II proporciona valores ilustrativos de coeficientes de regresión correspondientes a la biblioteca sistemáticamente variada ejemplificada en la Tabla I:

Tabla II: Ordenación por clasificación ilustrativa de los coeficientes de regresión

COEFICIENTE DE REGRESIÓN	VALOR
c_{166Ile}	62.15
c_{175Gly}	61.89
c_{10Asp}	60.23
c_{340Ala}	57.45
c_{10Ala}	50.12
c_{166Phe}	49.65
c_{166Leu}	49.42
c_{340Phe}	47.16
c_{166Ser}	45.34
c_{175Val}	43.65
c_{10Lys}	40.15

La lista ordenada por clasificación de coeficientes de regresión puede usarse para construir una nueva biblioteca de variantes de proteínas optimizadas con respecto a una actividad deseada (es decir, una idoneidad mejorada). Esto puede hacerse de varias maneras. Esto puede lograrse reteniendo los residuos de aminoácidos que tienen los coeficientes con los valores observados más altos. Estos son los residuos que el modelo de regresión indica que contribuyen más a la actividad deseada. Si se emplean descriptores negativos para identificar los residuos (por ejemplo, 1 para la leucina y -1 para la glicina), es necesario clasificar las posiciones de los residuos en función del valor absoluto del coeficiente. Obsérvese que, en tales situaciones, típicamente hay un único coeficiente para cada residuo. El valor absoluto de la magnitud del coeficiente da la clasificación de la posición del residuo correspondiente. A continuación, es necesario considerar los signos de los residuos individuales para determinar si cada uno de ellos es perjudicial o beneficioso en términos de la actividad deseada.

ii) Modelos no lineales

El modelado no lineal se emplea para tener en cuenta las interacciones residuo-residuo que contribuyen a la actividad de las proteínas. Un paisaje N-K describe este problema. El parámetro N se refiere al número de residuos variables en una colección de secuencias polipeptídicas relacionadas. El parámetro K representa la interacción entre residuos individuales dentro de cualquiera de estos polipéptidos. La interacción es habitualmente el resultado de la proximidad física entre varios residuos, ya sea en la estructura primaria, secundaria o terciaria del polipéptido. La interacción puede deberse a interacciones directas, interacciones indirectas, interacciones fisicoquímicas, interacciones debidas a productos intermedios de plegamiento, efectos traslacionales y similares. Consultar Kauffman, S. y Levin, S. (1987), "Towards a general theory of adaptive walks on rugged landscapes", Journal of Theoretical Biology 128 (1) 11-45.

El parámetro K se define de tal manera que para el valor K=1, cada residuo variable (por ejemplo, hay 20 de ellos) interactúa exactamente con otro residuo en su secuencia. En el caso de que todos los residuos estén física y químicamente separados de los efectos de todos los demás residuos, el valor de K es cero. Obviamente, dependiendo de la estructura del polipéptido, K puede tener una amplia variedad de valores diferentes. Con una estructura rigurosamente resuelta del polipéptido en cuestión, puede estimarse un valor de K. Sin embargo, a menudo no es este el caso.

Un modelo de actividad polipeptídica puramente lineal y aditivo (como el descrito anteriormente) puede mejorarse incluyendo uno o más términos de interacción no lineales que representen interacciones específicas entre 2 o más residuos. En el contexto de la forma de modelo presentada anteriormente, estos términos se representan como "productos cruzados" que contienen dos o más variables ficticias que representan los dos o más residuos particulares (cada uno asociado con una posición particular en la secuencia) que interactúan para tener un impacto positivo o negativo significativo sobre la actividad. Por ejemplo, un término de producto cruzado puede tener la forma $c_{ab}x_ax_b$, donde x_a es una variable ficticia que representa la presencia de un residuo particular en una posición particular de la secuencia y la variable x_b representa la presencia de un residuo particular en una posición diferente (que interactúa con la primera posición) en la secuencia polipeptídica. A continuación se muestra una forma de ejemplo detallada del modelo.

La presencia de todos los residuos representados en el término de producto cruzado (es decir, cada uno de los dos o más tipos específicos de residuos en posiciones específicamente identificadas) repercute sobre la actividad global del polipéptido. El impacto puede manifestarse de muchas maneras. Por ejemplo, cada uno de los residuos individuales que interactúan cuando están presentes solos en un polipéptido puede tener un impacto negativo en la actividad, pero cuando están presentes en el polipéptido, el efecto global es positivo. En otros casos puede producirse lo contrario. Además, puede producirse un efecto sinérgico, en el que cada uno de los residuos individuales por sí solo tiene un impacto relativamente limitado sobre la actividad, pero cuando todos ellos están presentes, el efecto sobre la actividad es mayor que los efectos acumulativos de todos los residuos individuales.

En algunas realizaciones, los modelos no lineales incluyen un término de producto cruzado para cada combinación posible de residuos variables que interactúan en la secuencia. Sin embargo, esto no representa la realidad física, ya que sólo un subconjunto de los residuos variables interactúan realmente entre sí. Además, se produciría un "sobreajuste" para producir un modelo que proporciona resultados espurios que son manifestaciones de los polipéptidos particulares usados para crear el modelo y no representan interacciones reales dentro del polipéptido. El número correcto de términos de producto cruzado para un modelo que represente la realidad física y evite el sobreajuste viene dictado por el valor de K. Por ejemplo, si K=1, el número de términos de interacción de producto cruzado es igual a N.

Al construir un modelo no lineal, en algunas realizaciones es importante identificar los términos de interacción de productos cruzados que representan verdaderas interacciones estructurales que tienen un impacto significativo sobre la actividad. Esto puede lograrse de varias maneras, incluyendo pero no limitadas a la adición directa, en la que los términos candidatos de productos cruzados se añaden al modelo inicial de sólo términos lineales de uno en uno hasta que la adición de términos ya no es estadísticamente significativa, y la sustracción inversa, en la que todos los posibles términos de productos cruzados se proporcionan en un modelo inicial y se eliminan de uno en uno. Los ejemplos ilustrativos que se presentan a continuación implican el uso de técnicas de adición y sustracción por pasos para identificar los términos de interacción no lineales útiles.

En algunas realizaciones, el enfoque para generar un modelo no lineal que contenga dichos términos de interacción es el mismo que el descrito anteriormente para generar un modelo lineal. En otras palabras, se emplea un conjunto de entrenamiento para "ajustar" los datos a un modelo. Sin embargo, se añaden al modelo uno o más términos no lineales, preferiblemente los términos de producto cruzado descritos anteriormente. Además, el modelo no lineal resultante, al igual que los modelos lineales descritos anteriormente, puede emplearse para clasificar la importancia de varios residuos sobre la actividad global de un polipéptido. Pueden usarse varias técnicas para identificar la mejor combinación de residuos variables según lo predicho por la ecuación no lineal. A continuación se describen los

enfoques para clasificar los residuos. En algunas realizaciones, se usa un gran número de posibles términos de productos cruzados para los residuos variables, incluso cuando se limitan a interacciones provocadas por sólo dos residuos. A medida que se producen más interacciones, el número de interacciones potenciales a considerar para un modelo no lineal crece de manera exponencial. Si el modelo incluye la posibilidad de interacciones que incluyan tres o más residuos, el número de términos potenciales crece aún más rápidamente.

En un ejemplo ilustrativo sencillo, en el que hay 20 residuos variables y K=1 (esto supone que cada residuo variable interactúa con otro residuo variable), debería haber 20 términos de interacción (productos cruzados) en el modelo. Si hay menos términos de interacción, el modelo no describirá completamente las interacciones (aunque algunas de las interacciones pueden no tener un impacto significativo sobre la actividad). Por el contrario, si hay más términos de interacción, el modelo puede sobreajustarse al conjunto de datos. En este ejemplo, hay N*(N-1)/2 o 190 posibles pares de interacciones. Encontrar la combinación de 20 pares únicos que describan las 20 interacciones de la secuencia es un problema computacional importante, ya que hay aproximadamente $5{,}48\times10^{26}$ combinaciones posibles.

Pueden emplearse numerosas técnicas para identificar los términos relevantes de los productos cruzados. Dependiendo del tamaño del problema y de la potencia de cálculo disponible, es posible explorar todas las combinaciones posibles e identificar de este modo el modelo que mejor se ajusta a los datos. Sin embargo, a menudo el problema es exigente desde el punto de vista computacional. Por tanto, en algunas realizaciones, se utiliza un algoritmo de búsqueda eficiente o una aproximación. Como se indica en la presente, una técnica de búsqueda adecuada es una técnica paso a paso. Sin embargo, no se pretende que la presente invención se limite a ningún método en particular para la identificación de los términos relevantes de los productos cruzados.

A continuación, en la Tabla III, se presenta un ejemplo ilustrativo para mostrar el valor de incorporar términos de productos cruzados no lineales en un modelo que predice la actividad a partir de la información de la secuencia. Este ejemplo es un modelo no lineal en el que se supone que sólo hay dos opciones de residuos en cada posición variable de la secuencia. En este ejemplo, la secuencia de la proteína se convierte en una secuencia codificada usando variables ficticias que corresponden a la opción A o a la opción B, usando +1 y -1 respectivamente. El modelo es inmune a la elección arbitraria de qué valor numérico se usa para asignar cada elección de residuo. Las posiciones de las variables que aparecen en la primera fila de la Tabla III no indican las posiciones reales de la secuencia de una proteína. En su lugar, son etiquetas arbitrarias que representan cualquiera de 10 posiciones hipotéticas en una secuencia de proteína que pueden variarse con una de las dos opciones mostradas en la segunda y tercera filas de la Tabla III para la Elección de Residuo A y la Elección de Residuo B.

TABLA III: Ejemplo de residuos codificantes en posiciones que tienen dos opciones cada una

Etiqueta de posición variable	1	2	3	4	5	6	7	8	9	10
Elección de Residuo A	I	L	L	M	G	W	K	C	S	F
Elección de Residuo B	V	A	I	P	H	N	R	T	A	Y
Elección del residuo de proteína	V	A	L	P	G	W	K	T	S	F
Modelo Código Valor	-1	-1	1	-1	1	1	1	-1	1	1

Con este esquema de codificación, el modelo lineal usado para asociar las secuencias de proteínas con la actividad puede escribirse de la siguiente manera:

$$y = c_1x_1 + c_2x_2 + c_3x_3 \ldots + c_nx_n + \ldots + c_Nx_N + c_0 \qquad \text{(Ec. 3)}$$

donde y es la respuesta (actividad), c_n el coeficiente de regresión para la elección del residuo en la posición n, x la variable ficticia que codifica la elección del residuo (+1/-1) en la posición n, y c_0 el valor medio de la respuesta. Esta forma del modelo asume que no hay interacciones entre los residuos variables (es decir, cada elección de residuo contribuye independientemente a la aptitud global de la proteína).

El modelo no lineal incluye un cierto número de términos de productos cruzados (aún por determinar) para tener en cuenta las interacciones entre residuos:

$$y = c_1x_1 + c_2x_2 + c_3x_3 + \ldots + c_nx_n + c_{1,2}x_1x_2 + c_{1,3}x_1x_3 + c_{2,3}x_2x_3 + \ldots + c_0 \qquad \text{(Ec. 4)}$$

donde las variables son las mismas que las de la Ec. (3) pero ahora hay términos no lineales, por ejemplo, $c_{1,2}$ es el coeficiente de regresión para la interacción entre las posiciones variables 1 y 2.

Para evaluar el rendimiento de los modelos lineales y no lineales, se usó una fuente de datos sintéticos conocida como paisaje NK (Kauffman y Levin, 1987). Como se ha mencionado anteriormente, N es el número de posiciones variables en una proteína simulada y K es el acoplamiento epistático entre residuos. Además, el conjunto de datos sintéticos se generó in silico.

Este conjunto de datos se usó para generar un conjunto de entrenamiento inicial con S=40 muestras sintéticas, con N=20 posiciones variables y K=1 (para reiterar, para K=1 cada posición variable está acoplada funcionalmente a otra posición variable). Al generar las proteínas aleatorias, cada posición variable tenía la misma probabilidad de contener la variable ficticia +1 o -1. Las interacciones residuo-residuo (representadas por productos cruzados) y las actividades reales se conocían para cada miembro del conjunto sintético de entrenamiento. Se generaron otras V=100 muestras para su uso en la validación. De nuevo, para cada miembro del conjunto de validación se conocían las interacciones y las actividades residuo-residuo.

Los conjuntos de entrenamiento se usaron para construir modelos lineales y no lineales. Algunos modelos no lineales se generaron con selección de los términos de productos cruzados y otros modelos no lineales se generaron sin selección de tales términos. Los modelos de las Figura 3A-F se generaron usando un método de modelado de algoritmo genético, mientras que los modelos de las Figura G-H se generaron usando el método de modelado por pasos. Aunque la ventaja cuantitativa de los modelos que tienen términos lineales y no lineales con respecto a los modelos que sólo tienen términos lineales difiere entre el algoritmo genético y los métodos de modelado por pasos, los resultados indican la ventaja generalizable de los modelos con términos no lineales, independientemente de los métodos de modelado. De hecho, no se pretende que la presente invención se limite a ningún método en particular, ya que en la presente invención puede usarse cualquier método de modelado adecuado.

Para el tamaño del conjunto de entrenamiento de S=40 descrito anteriormente, el modelo lineal fue capaz de correlacionar los valores medidos y predichos razonablemente bien, pero demostró una correlación más débil cuando se validó frente a datos no observados en el conjunto de entrenamiento (ver la Figura 3A). Como se muestra, los puntos de datos oscuros representan la actividad observada de 40 puntos de datos de entrenamiento frente a las predicciones realizadas por un modelo lineal. Los puntos de datos claros representan las predicciones realizadas por el mismo modelo construido a partir de las 40 muestras de entrenamiento y usado para predecir las V muestras de validación, ninguna de las cuales se observó en el conjunto de entrenamiento original. El conjunto de validación proporciona una buena medida de la verdadera capacidad de predicción del modelo, a diferencia del conjunto de entrenamiento, que puede adolecer del problema del sobreajuste del modelo, especialmente en los casos no lineales que se describen a continuación.

Este resultado para el conjunto de entrenamiento S=40 descrito anteriormente es notable, teniendo en cuenta que se usó un modelo lineal para modelar un paisaje de aptitud no lineal. En este caso, el modelo lineal podría, en el mejor de los casos, captar la contribución media a la aptitud para la elección de un residuo determinado. Si se tiene en cuenta un número suficiente de contribuciones medias combinadas, el modelo lineal predice aproximadamente la respuesta real medida. Los resultados de validación del modelo lineal fueron ligeramente mejores cuando se aumentó el tamaño del entrenamiento a S=100 (ver la Figura 3B). La tendencia de los modelos relativamente simples a infraajustar los datos se conoce como "sesgo".

Cuando el modelo no lineal se entrenó usando sólo S=40 muestras, la correlación con los miembros del conjunto de entrenamiento fue excelente (ver la Figura 3C). Desafortunadamente, en este ejemplo ilustrativo, el modelo proporcionó una potencia predictiva limitada fuera del conjunto de entrenamiento, como lo demuestra su correlación limitada con los valores medidos en el conjunto de validación. Este modelo no lineal, con muchas variables potenciales (210 posibles), y datos de entrenamiento limitados para facilitar la identificación de los términos de productos cruzados adecuados, fue capaz esencialmente de memorizar el conjunto de datos en el que fue entrenado. Esta tendencia de los modelos de alta complejidad a sobreajustar los datos se conoce como "varianza". El equilibrio entre sesgo y varianza representa un problema fundamental en el aprendizaje automático y casi siempre se requiere alguna forma de validación para abordarlo cuando se trata de problemas de aprendizaje automático nuevos o no caracterizados.

Sin embargo, cuando se entrenó el modelo no lineal usando un conjunto de entrenamiento mayor (S=100), como se muestra en la Figura 3D, el modelo no lineal funcionó extraordinariamente bien tanto para la predicción de entrenamiento como, lo que es más importante, para la predicción de validación. Las predicciones de validación fueron lo suficientemente precisas como para que la mayoría de los puntos de datos quedaran ocultos por los círculos oscuros usados para representar el conjunto de entrenamiento.

A modo de comparación, las Figuras 3E y 3F muestran el rendimiento de los modelos no lineales preparados sin una selección cuidadosa de los términos de los productos cruzados. A diferencia de los modelos de las Figuras 3C y 3D, se eligieron todos los términos de productos cruzados posibles (es decir, 190 términos de productos cruzados para N=20). Como se muestra en estas Figuras, la capacidad de predecir la actividad del conjunto de validación es relativamente pobre en comparación con la de los modelos no lineales generados con una cuidadosa selección de términos de productos cruzados. Esta pobre capacidad para predecir los datos de validación es una manifestación de sobreajuste.

Las Figuras 3G y 3H muestran respectivamente la potencia predictiva indicada por los residuos de un modelo lineal y un modelo no lineal por pasos para datos simulados in silico. El modelo no lineal por pasos se implementó como se describe en general más arriba y más específicamente a continuación.

Para probar estos modelos, se crearon datos simulados. Se creó un generador de números aleatorios ***R*** basado en una distribución normal con una media ***MN*** y una desviación típica ***SD***. A continuación, se definió un conjunto de 10 mutaciones. Se denominaron M1, M2... M10 (esta nomenclatura es arbitraria). Este paso simula la creación de diversidad

Cada mutación representaba un cambio de aminoácido en una posición dada dentro de una secuencia de proteína, y cada posición es independiente de las demás. A cada mutación anterior se le asignó un valor de actividad aleatorio ***A*** basado en ***R*** (***MN***=0, ***SD***=0,2). Se eligieron seis mutaciones anteriores y se emparejaron en tres pares ***P***. Estos pares representaban interacciones epistáticas entre mutaciones.

Se asignó un valor de actividad ***AP*** a cada par ***P*** basado en ***R*** (***MN***=0, ***SD***=0,2). Se construyó una biblioteca ***L*** de 50 variantes en la que cada variante contenía un número aleatorio de mutaciones ***M*** definido anteriormente; el número aleatorio de mutaciones se definió por el valor absoluto redondeado de **R** (***MN***=4, ***S***=0,25). Este paso simula la construcción y la secuenciación de la biblioteca.

La actividad de cada variante en ***L*** se calculó sumando primero a 1,0 (una actividad definida de la secuencia de tipo salvaje sin mutación) el valor de la actividad de cada mutación ***PA*** por pares (si ambas mutaciones estaban presentes), seguido de la adición de los valores de las mutaciones individuales restantes (***A***). El ruido del ensayo se simuló añadiendo al valor final de cada variante un valor aleatorio de ***R*** (***MN***=0, ***SD***=0,005). Este paso simula el cribado de variantes.

Se construyó un modelo lineal ***LM*** basado en los datos del último paso. Este modelo contenía diez variables/coeficientes independientes, cada uno de los cuales representaba una mutación de M. A continuación se ajustó el modelo lineal usando la regresión por mínimos cuadrados ordinarios y los datos obtenidos anteriormente.

Se usó un método de adición por pasos para seleccionar un modelo ***MM*** sobre la base de los datos obtenidos anteriormente, siendo el modelo base el ***LM***, usando el AIC como criterio de selección, y seleccionando modelos que sólo contienen coeficientes que representen mutaciones únicas e interacciones por pares. Para más detalles sobre el método de selección de modelos, consultar más adelante la descripción de la selección de modelos. El mejor modelo seleccionado por AIC se ajustó mediante regresión por mínimos cuadrados ordinarios.

Para evaluar la capacidad de predicción del modelo lineal y del modelo no lineal, los procedimientos descritos anteriormente se repitieron 20 veces. La predicción de los modelos se trazó contra los datos simulados, y en la Figura 3G se muestra el modelo lineal y en la Figura 3H el modelo no lineal por pasos. Los modelos se usaron para predecir los valores de las mutaciones individuales descritas anteriormente. Esta predicción se realizó usando los modelos para predecir una variante que contenía una sola mutación de interés y restando 1,0 (tipo salvaje). Como resulta evidente de las Figuras 3G y 3H, el modelo no lineal predice con mayor precisión los valores, con una tendencia más lineal y residuos más pequeños.

iii) Selección del modelo

Se usan métodos de adición o sustracción por pasos para preparar modelos con términos de interacción no lineales. Al implementar la operación mostrada en el bloque 107 de la Figura 1, se proporciona un modelo final con alta potencia predictiva que incluye términos de interacción mediante la adición o sustracción por pasos de términos de interacción de un modelo base. La Figura 4A proporciona un diagrama de flujo de una implementación de la operación del bloque 107 de la Figura 1 añadiendo términos de interacción a un modelo base y evaluando los nuevos modelos para crear un mejor modelo final.

En este ejemplo, el modelo de secuencia base no incluye términos de interacción. El método establece primero un modelo de secuencia actual y un modelo de secuencia mejor en el modelo de secuencia base, bloque 409. El método define un conjunto de términos de interacción para las variantes de secuencia. Estos términos de interacción pueden incluir cualquier número de interacciones por pares o de orden superior de dos o más residuos de aminoácidos. Ver el bloque 411. Aunque el bloque 409 se ilustra como anterior al bloque 411, el orden de los dos pasos no es importante. En algunas realizaciones, el conjunto de términos de interacción incluye combinaciones factoriales de todos los residuos de aminoácidos de interés. En algunas realizaciones adicionales, se incluyen por lo menos todos los términos de interacción por pares. En algunas realizaciones adicionales, se incluyen términos de interacción por pares y de tres vías.

Después de crear un modelo base, el método selecciona un término de interacción del conjunto que aún no se ha probado. A continuación, el método crea un nuevo modelo de secuencia añadiendo el término de interacción seleccionado al modelo de secuencia actual. Ver el bloque 413. A continuación, el método evalúa la potencia predictiva del nuevo modelo de secuencia usando un método de selección de modelos con un sesgo en contra de la inclusión de términos de interacción adicionales. Ver el bloque 415. El método determina si la potencia predictiva del nuevo modelo de secuencia es mayor o no que la del mejor modelo de secuencia. Ver el bloque de decisión 417. A modo de

ejemplo, el método puede usar una técnica que emplee la determinación de la "verosimilitud" (por ejemplo, AIC) como criterio de selección del modelo. En tales casos, sólo se considera que tiene mayor potencia predictiva un modelo que tenga un valor de AIC menor que el modelo previamente probado.

El método de selección se sesga contra de los modelos con más parámetros. Ejemplos de tales métodos de selección incluyen, pero no se limitan a, el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC), y sus variaciones. Por ejemplo, AIC puede calcularse como:

$$AIC = -2log_eL + 2k$$

donde L es la verosimilitud de un modelo dado un conjunto de datos, y k es el número de parámetros libres en un modelo.

En algunas realizaciones, la verosimilitud de un modelo dado un conjunto de datos puede calcularse mediante varios métodos, incluyendo, entre otros, el método de máxima verosimilitud. Por ejemplo, para una variable dependiente binaria en la que una actividad está presente o ausente en una observación, la probabilidad del modelo puede calcularse como:

$$L(modeloldatos) = \prod_{i=1}^{n} \frac{(a_i + b_i)!}{a_i!\, b_i!} p_i^{a_i} (1 - p_i)^{b_i}$$

en donde n es el número total de puntos de datos en un conjunto de datos; a_i y b_i son el número de ensayos observados que comprenden la condición *i-ésima*; p es la probabilidad de que una variable dependiente se observe según lo predicho por el modelo.

En algunas realizaciones que implican una serie de modelos anidados, como en los modelos de regresión con un número progresivamente mayor de términos de interacción (y coeficientes asociados) que un modelo base, los modelos más complejos proporcionan ajustes igual de buenos o mejores que los más sencillos, incluso si los coeficientes adicionales son espurios, porque el modelo más complejo disfruta de grados de libertad adicionales. En algunas realizaciones, el AIC penaliza el modelo más complejo en la medida en que la ganancia en la bondad del ajuste se ve más que compensada por el coste de los parámetros espurios. En la selección de modelos, un valor menor de AIC indica un modelo mejor.

En el ejemplo mostrado en la Figura 4A, si la potencia predictiva del nuevo modelo de secuencia es mayor que el del mejor modelo de secuencia, entonces el método establece el nuevo modelo como el mejor modelo. Ver el bloque 419. A continuación, el método comprueba si quedan términos de interacción adicionales en el conjunto que no se han probado. Ver el bloque de decisión 421. Si es así, el proceso vuelve al bloque 413, formando de este modo un bucle interno para probar todos los términos de interacción disponibles en el conjunto de interacciones. A través de iteraciones del bucle interno, puede encontrarse el mejor término de interacción y añadirlo al modelo.

Una vez probados todos los términos de interacción y finalizado el bucle interno, se identifica un mejor modelo con un término de interacción adicional, dado que existe un modelo con mayor potencia predictiva que el mejor modelo anterior. Ver el bloque de decisión 423. En tales realizaciones, el método establece el modelo actual como el mejor modelo y excluye los términos de interacción del mejor modelo del conjunto disponible de términos de interacción. Ver el bloque 425. A continuación, el método vuelve al bloque 413. Este bucle externo busca el siguiente mejor término de interacción que pueda mejorar la potencia predictiva del modelo. Si se encuentra tal término de interacción, la búsqueda del siguiente mejor término de interacción continúa en el bucle exterior, hasta que no se identifique ningún modelo nuevo que tenga una potencia predictiva mayor que la del mejor modelo de secuencia anterior.

Cuando no se pueden encontrar más términos de interacción para mejorar el modelo, el método establece el mejor modelo como modelo final. Ver el bloque 427. La búsqueda del mejor modelo dados los datos de secuencia y actividad ha terminado. El modelo se usa entonces para predecir las actividades de nuevas secuencias. Tales predicciones pueden guiar la selección de secuencias para su posterior variación y prueba.

En ciertos métodos, cada uno de los términos de interacción disponibles en el grupo de términos de interacción se trata como si tuviera potencialmente el mismo impacto en la calidad o la potencia predictiva del modelo. En otras palabras, en la implementación, cada uno de los términos de interacción disponibles en el grupo tiene la misma probabilidad de ser seleccionado para su consideración durante una iteración particular. Como se define en las reivindicaciones, los términos de interacción disponibles se seleccionan aleatoriamente o en un orden arbitrario. En algunos otros métodos, los términos de interacción están sesgados o ponderados de tal manera que es más probable que algunos términos sean seleccionados para su consideración que otros durante una iteración dada. El sesgo o la ponderación pueden, en ciertos métodos, aplicarse sobre la base de información física o teórica acerca de las interacciones. Por ejemplo, puede saberse que es probable que las mutaciones en dos áreas particulares de una

proteína estén físicamente próximas entre sí y, por tanto, interactúen. Los términos de interacción pertenecientes a los residuos de estas dos áreas generales podrían estar sesgados para su selección durante el proceso iterativo de refinamiento del modelo.

A continuación se muestra un pseudocódigo que ilustra procesos similares al de la Figura 4A:

```
SET Coeff = Interaction Terms to Test
Best = Baseline Model
count = 1
WHILE count > 0
        count = 0
        BestFromRound = Best
         BestCoefficient = NULL
        FOR each Interaction Term in Coeff
                TestModel = (best + Interaction Term)[1]
                IF TestModel BETTER THAN BestFromRound  THEN[2]
                        BestFromRound = TestModel
                        Count++
                        BestCoefficient = Interaction Term
                ENDIF
        ENDFOR
        IF count > 0 THEN
                Best = BestFromRound
                Remove BestCoefficient FROM Coeff[3]
        ENDIF
ENDWHILE
```

El elemento 1 añade el término de interacción de la prueba al modelo de regresión

El elemento 2 representa la Comparación de Modelos, uno o más de los Criterios de Información de Akaike (AIC), Criterios de Información Bayesiana (BIC), Validación Cruzada (error medio), ANOVA, o contribución de coeficientes.

El elemento 3 se proporciona para evitar duplicar las pruebas de términos de interacción.

La Figura 4B proporciona un diagrama de flujo que muestra una realización de la operación mostrada en el bloque 107 de la Figura 1. En este proceso, los términos de interacción se restan de un modelo base que incluye todos los términos de interacción posibles de un grupo de tales términos para crear un mejor modelo final.

En esta realización, el modelo de secuencia base incluye todos los términos de interacción dentro de un grupo definido. El método establece primero un modelo de secuencia actual y un modelo de secuencia mejor para que sean iguales al modelo de secuencia base al comienzo del proceso, bloque 439. Esta realización es similar al último modelo descrito anteriormente en que el grupo de términos de interacción puede incluir cualquier número de interacciones por pares o de orden superior de dos o más residuos de aminoácidos. En algunas realizaciones, el grupo de términos de interacción incluye combinaciones factoriales de todos los residuos de aminoácidos que son de interés.

Después de crear un modelo base, el método selecciona un término de interacción que aún no se ha probado del grupo de términos ya incluidos en el modelo base. A continuación, el método crea un nuevo modelo de secuencia restando el término de interacción seleccionado del modelo de secuencia actual. Ver el bloque 441. A continuación, el método evalúa la potencia predictiva del nuevo modelo de secuencia usando un método de selección de modelos con un sesgo en contra de términos de interacción adicionales. Ver el bloque 443. El método evalúa si la potencia predictiva del nuevo modelo de secuencia es mayor o no que el del mejor modelo de secuencia. Ver la operación de decisión mostrada en el bloque 445. En algunas realizaciones, se usa el AIC como criterio de selección del modelo, de forma que un modelo que tenga un valor de AIC menor que el modelo probado previamente se considera que tiene mayor potencia predictiva.

En este ejemplo ilustrativo, si la potencia predictiva del nuevo modelo de secuencia es mayor que la del mejor modelo de secuencia, entonces el método establece el nuevo modelo como el mejor modelo. Ver el bloque 447. A continuación, el método comprueba si quedan términos de interacción adicionales en el grupo que no se han probado (es decir, restados del modelo de secuencia actual). Ver el bloque de decisión 449. Si hay términos sin probar, el

método vuelve al bloque 441, formando de este modo un bucle interno para probar todos los términos de interacción disponibles en el grupo de interacciones. A través de iteraciones del bucle interno, se identifica un único término de interacción. Al eliminarlo del modelo se mejora el modelo en la mayor medida (y se reduce el AIC en el mayor margen, si se usa el AIC para medir la potencia predictiva del modelo).

Una vez probados todos los términos de interacción y finalizado el bucle interno, se identifica un mejor modelo que tenga un término de interacción menos, dado que existe un modelo que tiene mayor potencia predictiva que el mejor modelo anterior. Ver el bloque de decisión 451. En este caso, el método establece el modelo actual como el mejor modelo. Ver el bloque 453. A continuación, el método vuelve al bloque 441. Este bucle externo busca el siguiente término de interacción que pueda mejorar la potencia predictiva del modelo con el mayor margen. Si se encuentra tal término de interacción, la búsqueda del siguiente término de interacción a sustraer continúa en el bucle exterior, hasta que no se identifiquen más modelos nuevos que tengan potencias predictivas mayores que la del mejor modelo de secuencia anterior.

Cuando se completa un bucle interno y no pueden encontrarse más términos de interacción que sustraer para mejorar el modelo (es decir, la operación de decisión mostrada en el bloque 451 se responde negativamente), el método establece el último mejor modelo como modelo final. Ver el bloque 455. La búsqueda del mejor modelo dados los datos de secuencia y actividad ha terminado.

iv) Opciones alternativas de modelización

Dentro del alcance de la divulgación se incluyen múltiples variaciones adicionales del enfoque anterior. De hecho, no se pretende que la presente invención se limite a ningún modelo en particular, ya que en la presente invención encuentra uso cualquier modelo adecuado. Como ejemplo ilustrativo, las variables x_{ij} son representaciones de las propiedades físicas o químicas de los aminoácidos, en lugar de las identidades exactas de los propios aminoácidos (leucina frente a valina frente a prolina...). Algunos ejemplos de estas propiedades son la lipofilia, el volumen y las propiedades electrónicas (por ejemplo, la carga formal, la superficie de Van der Waals asociada a una carga parcial, etc.). Para implementar este enfoque, los valores x_{ij} que representan residuos de aminoácidos pueden presentarse en términos de sus propiedades o componentes principales construidos a partir de estas propiedades. No se pretende que la presente invención se limite a ninguna propiedad particular de aminoácidos, péptidos y/o polipéptidos, ya que cualquier propiedad adecuada encuentra uso en los métodos de la presente invención.

En algunas realizaciones adicionales, las variables x_{ij} representan nucleótidos, en lugar de residuos de aminoácidos. En estas realizaciones, el objetivo es identificar secuencias de ácido nucleico que codifican proteínas para una biblioteca de variantes de proteínas. Al usar nucleótidos en lugar de aminoácidos, pueden optimizarse parámetros distintos de la actividad (por ejemplo, la actividad específica), según se desee. Por ejemplo, la expresión de la proteína en un huésped o vector particulares puede ser una función de la secuencia de nucleótidos. Dos secuencias de nucleótidos diferentes pueden codificar una proteína que tenga la misma secuencia de aminoácidos, pero una de las secuencias de nucleótidos puede llevar a la producción de mayores cantidades de proteína y/o la proteína es más activa. Al usar secuencias de nucleótidos en lugar de secuencias de aminoácidos, los métodos descritos en la presente pueden emplearse para optimizar cepas de microorganismos que presenten mejores propiedades de expresión génica y/o mejores propiedades (por ejemplo, actividad específica, estabilidad, etc.).

En algunas realizaciones, la secuencia de nucleótidos se representa como una secuencia de codones. En algunas realizaciones, los modelos utilizan codones como la unidad atómica de una secuencia de nucleótidos, de tal manera que las actividades previstas son una función de los varios codones presentes en la secuencia de nucleótidos. Cada codón, junto con su posición en la secuencia global de nucleótidos, sirve como variable independiente para generar modelos de secuencia-actividad. Se observa que, en algunos casos, diferentes codones para un aminoácido determinado se expresan de manera diferente en un organismo determinado. En algunas realizaciones, cada organismo tiene un codón preferido, o una distribución de frecuencias de codones, para un aminoácido dado. Al usar codones como variables independientes, la realización tiene en cuenta estas preferencias. Por tanto, la realización puede usarse para generar una biblioteca de variantes de expresión (por ejemplo, donde "actividad" incluye el nivel de expresión génica de un organismo huésped particular).

En algunas realizaciones, los métodos incluyen las siguientes operaciones: (a) recibir datos que caracterizan un conjunto de entrenamiento de una biblioteca de variantes de proteínas; (b) desarrollar un modelo secuencia-actividad no lineal que predice la actividad como una función de los tipos de nucleótidos y las posiciones correspondientes en la secuencia de nucleótidos, basándose en los datos obtenidos en (a); (c) usar el modelo de secuencia-actividad para clasificar las posiciones en una secuencia de nucleótidos y/o tipos de nucleótidos en posiciones específicas en la secuencia de nucleótidos en orden de impacto sobre la actividad deseada; y (d) usar la clasificación para identificar uno o más nucleótidos, en la secuencia de nucleótidos, que deben variarse o fijarse, con para mejorar la actividad deseada. Como se ha indicado, en algunas realizaciones, los nucleótidos a variar codifican aminoácidos específicos.

En algunos otros ejemplos, los métodos implican el uso de diferentes técnicas para clasificar o caracterizar

de otro modo los residuos en términos de su importancia con respecto a una determinada propiedad. Como se describió anteriormente para los modelos lineales, se usaron las magnitudes de los coeficientes de regresión para clasificar los residuos. Los residuos que tenían coeficientes con magnitudes grandes (por ejemplo, 166 Ile) se consideraron residuos de alta clasificación. Esta caracterización se usó para decidir si variar o no un residuo concreto en la generación de una nueva biblioteca optimizada de variantes de proteínas. Para los modelos no lineales, el análisis de sensibilidad era más complejo, como se describe en la presente.

El PLS y otras técnicas proporcionan información adicional, más allá de la magnitud del coeficiente de regresión, que puede usarse para clasificar residuos específicos o posiciones de residuos. Técnicas como PLS y el Análisis de Componentes Principales (ACP) o PCR proporcionan información en forma de componentes principales o vectores latentes. Éstos representan direcciones o vectores de variación máxima a través de conjuntos de datos multidimensionales como el espacio de secuencia-actividad de proteínas empleado con las realizaciones de la presente invención divulgadas en la presente. Estos vectores latentes son funciones de las varias dimensiones de secuencia; es decir, los residuos individuales o posiciones de residuos que comprenden las secuencias de proteínas que comprenden la biblioteca de variantes usada para construir el conjunto de entrenamiento. Por lo tanto, los vectores latentes comprenden una suma de las contribuciones de cada una de las posiciones de residuos en el conjunto de entrenamiento. Algunas posiciones contribuyen en mayor medida a la dirección del vector. Éstas se manifiestan mediante "cargas" relativamente grandes, es decir, los coeficientes usados para describir el vector. Como ejemplo ilustrativo sencillo, un conjunto de entrenamiento puede estar compuesto por tripéptidos. En este ejemplo, el primer vector latente comprende contribuciones de los tres residuos.

Vector 1 = a1(posición del residuo 1) + a2(posición del residuo 2) + a3(posición del residuo 3)

Los coeficientes a1, a2 y a3 son las cargas. Debido a que reflejan la importancia de las posiciones de residuos correspondientes para la variación en el conjunto de datos, pueden usarse para clasificar la importancia de las posiciones de residuos individuales a efectos de las decisiones de "alternancia", como se ha descrito anteriormente. Las cargas, al igual que los coeficientes de regresión, pueden usarse para clasificar los residuos en cada posición alternada. Varios parámetros describen la importancia de estas cargas. Algunas realizaciones utilizan métodos como la Importancia Variable en la Proyección (VIP) para hacer uso de una matriz de carga. Esta matriz de carga se compone de las cargas para múltiples vectores latentes tomados de un conjunto de entrenamiento. En los métodos de Importancia Variable para la Proyección PLS, la importancia de una variable (por ejemplo, la posición del residuo) se computa calculando VIP. Para una dimensión de PLS dada, a, $(VIN)_{ak}^2$ es igual al peso PLS al cuadrado $(w)_{ak}^2$ de una variable multiplicado por el porcentaje de variabilidad explicada en y (variable dependiente, por ejemplo, cierta función) por esa dimensión PLS. $(VIN)_{ak}^2$ se suma sobre todas las dimensiones PLS (componentes). A continuación, se calcula el VIP dividiendo la suma por el porcentaje total de variabilidad de y explicado por el modelo PLS y multiplicándolo por el número de variables del modelo. Las variables con un VIP superior a 1 son las más relevantes para correlacionarse con una determinada función (y) y, por tanto, las mejor clasificadas para tomar decisiones de alternancia.

En muchas realizaciones, la presente invención utiliza métodos de regresión lineal general para identificar los efectos de las mutaciones en una biblioteca combinatoria sobre una secuencia-actividad de interés. Pueden usarse opciones y técnicas de modelado alternativas, por ejemplo, regresión bayesiana, regresión de conjunto, bootstrapping, en combinación con o en lugar de los métodos mencionados anteriormente. De hecho, no se pretende que la presente invención se limite a ninguna opción y/o técnica de modelización específica, ya que en la presente invención puede usarse cualquier método o métodos adecuados.

Regresión lineal bayesiana

En algunas realizaciones de la presente invención, se usa la regresión lineal bayesiana. Este método es un enfoque de la regresión lineal en el que el análisis estadístico se lleva a cabo en el contexto de la inferencia bayesiana. Cuando el modelo de regresión tiene errores que presentan una distribución normal, y si se asume una forma particular de distribución a priori, las distribuciones de probabilidad a posteriori de los parámetros del modelo pueden determinarse usando técnicas de inferencia bayesiana.

Una solución de mínimos cuadrados ordinarios de un modelo de regresión lineal estima el vector de coeficientes y el error del modelo basándose en la función de verosimilitud de los datos mediante un método de cálculo analítico como la pseudoinversa de Moore-Penrose. Se trata de un enfoque frecuentista que supone que hay suficientes observaciones de los datos para representar la relación secuencia-actividad de todas las secuencias. Sin embargo, las observaciones reales de una muestra casi nunca son suficientes para representar a todos los miembros de una población. Esto es especialmente problemático cuando el tamaño de la muestra (o conjunto de entrenamiento) es limitado. En el enfoque bayesiano, los datos de la muestra se suplementan con información adicional en forma de distribución de probabilidad a priori. La creencia a priori sobre los parámetros se combina con la función de verosimilitud de los datos de acuerdo con el teorema de Bayes para obtener la creencia a posteriori sobre los parámetros. La creencia a priori puede adoptar distintas formas funcionales en función del dominio y de la información disponible a priori.

Por ejemplo, en algunas realizaciones, la regresión bayesiana puede usar información previa para ponderar los coeficientes antes del ajuste del modelo. En algunas realizaciones, para ponderar los coeficientes lineales pueden usarse los datos de secuencia/actividad tomados de una ronda previa de evolución dirigida, por ejemplo, una ronda realizada usando la estructura principal parental o de referencia y por lo menos algunas de las mutaciones usadas en las rondas previas. Además, pueden usarse las predicciones de la relación epistática entre dos o más mutaciones para ponderar coeficientes de interacción no lineales. Una de las principales ventajas de este enfoque es la inclusión de información previa para dirigir las predicciones del modelo.

Un ejemplo ilustrativo de una fuente de información previa es un modelo con términos independientes y de interacción para cada una de las múltiples mutaciones de una estructura principal de referencia. En algunas realizaciones, los datos se obtienen de una colección de variantes que contiene una mutación por variante.

Ejemplos adicionales de información previa que encuentran uso en la presente invención incluyen, pero no se limitan a, información intuitiva o física sobre el papel de ciertas mutaciones o tipos de mutaciones. Independientemente de la fuente, la información previa sirve como noción preconcebida de la relación entre secuencia y actividad.

En algunas realizaciones, para estimar los parámetros de un modelo, la regresión lineal bayesiana usa simulaciones de Monte Carlo, como el muestreo de Gibbs o los algoritmos de Metrópolis, para ajustar el modelo a los datos. El muestreo de Gibbs es un algoritmo de Monte Carlo de cadena de Markov para obtener una secuencia de observaciones que proceden aproximadamente de una distribución de probabilidad multivariante especificada (es decir, de la distribución de probabilidad conjunta de dos o más variables aleatorias), cuando es difícil el muestreo directo.

La Figura 5 es un diagrama de flujo que ilustra el uso de la regresión bayesiana en la evolución guiada de bibliotecas de variantes. Cada ronda de evolución de secuencias incluye mutaciones basadas en las secuencias de una ronda anterior, que pueden guiarse por conocimientos como un modelo de secuencia-actividad. En la ronda n de la evolución, como en el bloque 501, por ejemplo, hay una mutación por variante. La ronda siguiente o n+1 de la evolución es la ronda actual, como se muestra en el bloque 503. Hay por lo menos una nueva mutación para cada variante, lo que supone dos o más mutaciones por variante. E este ejemplo ilustrativo la regresión bayesiana se implementa en esta ronda.

Las variantes de secuencia de la ronda n+1 proporcionan un conjunto de datos de entrenamiento para nuevos modelos. Los nuevos modelos pueden comprender un modelo base que incluya sólo términos lineales para residuos individuales, o un modelo completo que contenga todos los posibles términos/coeficientes de interacción, como se indica en el bloque 507. Los nuevos modelos también pueden comprender un modelo seleccionado mediante varias técnicas, incluyendo las técnicas de adición o sustracción por pasos explicadas anteriormente, ver el bloque 505. Alternativamente, el modelo puede seleccionarse mediante un algoritmo genético o técnicas bootstrap, como se analiza más adelante. Todos estos modelos se basan en los datos actuales/nuevos del conjunto de datos de entrenamiento de la ronda n+1. A estos modelos puede aplicarse la técnica de inferencia bayesiana, de tal manera que un modelo se basa tanto en la función de probabilidad de los datos actuales como en la distribución de la información previa. La información previa puede proceder de los datos de la ronda anterior de variantes de secuencia, como en la ronda n indicada por el bloque 501. La información también puede proceder de los datos de secuencia-actividad de cualquier ronda anterior de evolución, o de otra intuición previa sobre el conocimiento, como se indica en el bloque 513. El modelo de regresión bayesiano indicado por el bloque 509 predice la actividad basándose en la información proporcionada por los datos actuales y la información previa, ver el bloque 511. Aunque la Figura 5 sólo ilustra la aplicación de la técnica de regresión bayesiana a la ronda n+1, puede aplicarse en varias etapas. Tampoco se pretende que la presente invención se limite a los pasos específicos proporcionados en la Figura 5, ya que en la presente invención encuentra uso cualquier método adecuado.

Regresión de conjunto

En algunas realizaciones, para preparar el modelo secuencia-actividad la presente invención utiliza una técnica de regresión de conjunto. Un modelo de regresión de conjunto se basa en varios modelos de regresión. La predicción de cada modelo se pondera basándose en un criterio de información (IC) particular, y la predicción del conjunto es una suma ponderada de la predicción de todos los modelos que contiene. En algunas realizaciones, el desarrollo del modelo comienza con un modelo base que contiene todos los términos lineales. Los modelos posteriores se construyen añadiendo coeficientes de interacción en alguna o todas las combinaciones posibles. En algunas realizaciones, los coeficientes de interacción se añaden en un proceso paso a paso. Cada modelo se ajusta a los datos y se genera un IC. La ponderación de cada modelo se basa en el IC, que puede ser el propio IC o una versión transformada, por ejemplo, un valor logarítmico, un valor negado, etc. Pueden hacerse predicciones para una observación generando la predicción de cada modelo del conjunto y determinando la predicción del conjunto tomando la media ponderada de la predicción de cada modelo. Un conjunto completo contiene todos los modelos posibles, pero puede recortarse para eliminar los modelos de bajo rendimiento estableciendo un umbral en el número de modelos que contiene o en el IC.

Los modelos constituyentes del conjunto pueden producirse usando varias técnicas. Por ejemplo, en algunas realizaciones, se usa un algoritmo genético para crear los modelos constituyentes. Los datos de secuencia/actividad se usan para producir una pluralidad de modelos de regresión, cada uno de los cuales tiene su propio conjunto de coeficientes. Los mejores modelos se seleccionan según un criterio de adecuación (por ejemplo, AIC o BIC). Estos modelos se "aparean" para producir nuevos modelos híbridos cuya idoneidad se evalúa y se selecciona en consecuencia. En algunas realizaciones, este proceso se repite en múltiples rondas de "evolución computacional" para producir un conjunto de los mejores modelos. Alternativamente, en algunas realizaciones, los componentes del conjunto se crean por regresión por pasos como se ha descrito anteriormente, y se seleccionan los n mejores modelos para formar un conjunto.

La Figura 6 proporciona un diagrama de flujo de un proceso que implementa la regresión de conjuntos en la evolución dirigida de variantes de secuencias de acuerdo con una realización de la presente invención. En esta realización, la técnica de regresión de conjunto puede aplicarse en cualquier etapa de múltiples rondas de evolución de secuencia. Por ejemplo, en la ronda n, las variantes de secuencia mostradas en el bloque 601 proporcionan un conjunto de datos de entrenamiento para varios modelos para formar un grupo de modelos como se indica en el bloque 603. Los modelos del grupo de modelos pueden ser modelos generados por un algoritmo genético y/o selección por pasos. En otras realizaciones, el grupo de modelos comprende modelos de validación cruzada n-veces y/o modelos bootstrapping. En algunas realizaciones, sólo se seleccionan los modelos con un potencia predictiva superior para entrar en el grupo basándose en varios criterios de selección de modelos, como AIC o BIC.

Alternativa o adicionalmente, en algunas realizaciones, los modelos que no han sido examinados por la selección de modelos también entran en la reserva de modelos. En una realización, se introducen en el conjunto de modelos todos los modelos con todos los términos lineales y no lineales. Para un gran número de residuos y un número mucho mayor de interacciones factoriales entre residuos, esta realización puede ser muy intensiva computacionalmente. En algunas realizaciones alternativas, sólo se introducen en el conjunto de modelos los modelos que contienen términos lineales y términos de interacción por pares. Independientemente del método de inclusión del grupo de modelos, un modelo de conjunto incluye todos los términos de sus constituyentes. El grupo de modelos puede contener cualquier número de modelos, incluidos, entre otros, los modelos bayesianos, en cuyo caso puede incorporarse información previa al conjunto.

En algunas realizaciones, el conjunto predice la actividad de la secuencia basándose en la media ponderada de los coeficientes de cada modelo del grupo, en donde las ponderaciones se determinan por la potencia predictiva de los modelos correspondientes, como se indica en el bloque 605.

En algunas realizaciones, una regresión de conjunto usa el siguiente flujo de trabajo: (1) proporcionar un conjunto vacío; (2) seleccionar un tamaño de grupo n de 1 o mayor; (3) categorizar los puntos de datos en grupos de tamaño n, donde los puntos de datos se agrupan sin reemplazo; y (4) preparar un modelo de conjunto para predecir los coeficientes individuales y de interacción. En algunas realizaciones, el paso (4) para preparar un modelo de conjunto comprende además: a) eliminar puntos de datos de cada grupo, en donde los datos restantes forman un conjunto de entrenamiento y los datos excluidos forman un conjunto de validación; b) preparar un modelo ajustando el conjunto de entrenamiento usando regresión por pasos; c) probar el modelo usando el conjunto de validación, que proporciona una indicación de la capacidad predictiva del modelo; d) añadir el modelo a un grupo de modelos que se usan para generar un modelo de conjunto como se ha descrito anteriormente.

Enfoque Bootstrap

En la presente invención encuentran uso otras técnicas para caracterizar la potencia predictiva de un modelo en consideración en una iteración dada. En algunas realizaciones, estas técnicas implican validación cruzada o técnicas Bootstrap. En algunas realizaciones, la validación cruzada emplea un conjunto de observaciones usadas para generar el modelo, pero deja algunas de las observaciones fuera para evaluar la fuerza del modelo. En algunas realizaciones, la técnica Bootstrap implica el uso de un conjunto de muestras que se prueban con reemplazo. En algunas realizaciones, los modelos generados mediante validación cruzada o Bootstrap pueden combinarse en un modelo de conjunto como se ha descrito anteriormente.

En algunas realizaciones adicionales, los métodos clasifican los residuos no sólo por las magnitudes de sus contribuciones previstas a la actividad, sino también por la confianza en esas contribuciones previstas. En algunos casos, al investigador le preocupa la generalizabilidad del modelo de un conjunto de datos a otro conjunto. En otras palabras, el investigador quiere saber si los valores de los coeficientes o componentes principales son espurios o no. Las técnicas de validación cruzada y bootstrapping proporcionan medidas para indicar el nivel de confianza en que los modelos son generalizables a varios datos.

En algunas realizaciones, se usa un enfoque estadísticamente más riguroso en el que la clasificación se basa en una combinación de magnitud y distribución. En algunas de estas realizaciones, los coeficientes con magnitudes elevadas y distribuciones ajustadas obtienen la clasificación más alta. En algunos casos, un coeficiente con una

magnitud menor que otro puede recibir una clasificación más alta en virtud de tener menos variación. Por tanto, algunas realizaciones clasifican los residuos de aminoácidos o nucleótidos basándose tanto en la magnitud como en la desviación estándar o varianza. Pueden usarse varias técnicas para lograr esto. De hecho, no se pretende que la presente invención se limite a ninguna técnica específica para la clasificación. A continuación se describe una realización que usa un enfoque de valor p de Bootstrap.

En la Figura 7 se muestra un ejemplo ilustrativo de un método que emplea un método Bootstrap. Como se muestra en la Figura 7, el método 725 comienza en el bloque 727, donde se proporciona un conjunto de datos original S. En algunas realizaciones, se trata de un conjunto de entrenamiento como se ha descrito anteriormente. Por ejemplo, en algunas realizaciones, se genera variando sistemáticamente los residuos individuales de una secuencia de partida de cualquier manera (por ejemplo, como se ha descrito anteriormente). En el caso ilustrado por el método 725, el conjunto de datos S tiene M puntos de datos diferentes (información de actividad y secuencia recogida de secuencias de aminoácidos o nucleótidos) para su uso en el análisis.

A partir del conjunto de datos S, se crean varios conjuntos de Bootstrap B. Cada uno de estos conjuntos se obtiene por muestreo, con reemplazo, del conjunto S para crear un nuevo conjunto de M miembros, todos tomados del conjunto original S. Ver el bloque 729. La condición "con reemplazo" produce variaciones del conjunto original S. El nuevo conjunto de Bootstrap, B, contendrá a veces muestras replicadas de S. La condición "con reemplazo" produce variaciones en el conjunto original S. El nuevo conjunto de Bootstrap, B, a veces contendrá muestras replicadas de S. En algunos casos, el conjunto de Bootstrap B también carece de ciertas muestras originalmente contenidas en S.

Como ejemplo ilustrativo, se proporciona un conjunto S de 100 secuencias. Se crea un conjunto de Bootstrap B seleccionando aleatoriamente 100 secuencias miembros de las 100 secuencias del conjunto original S. Cada conjunto Bootstrap B usado en el método contiene 100 secuencias. Por tanto, es posible que algunas secuencias se seleccionen más de una vez y que otras no se seleccionen en absoluto. Usando el conjunto de Bootstrap B producido a partir del conjunto S de 100 secuencias, el método construye a continuación un modelo. Ver el bloque 731. El modelo puede construirse como se ha descrito anteriormente, usando PLS, PCR, una SVM, regresión por pasos, etc. De hecho, se pretende que en la construcción del modelo encuentre uso cualquier método adecuado. Este modelo proporciona coeficientes u otros indicios de clasificación para los residuos o nucleótidos encontrados en las varias muestras del conjunto B. Como se muestra en el bloque 733, se registran estos coeficientes u otros indicios para su uso posterior.

A continuación, en el bloque de decisión 735, el método determina si debe crearse otro conjunto de Bootstrap. En caso afirmativo, el método vuelve al bloque 729 donde se crea un nuevo conjunto de Bootstrap B como se ha descrito anteriormente. En caso negativo, el método pasa al bloque 737 que se describe a continuación. La decisión en el bloque 735 depende de cuántos conjuntos diferentes de valores de coeficientes se usarán para evaluar las distribuciones de esos valores. El número de conjuntos B debe ser suficiente para generar estadísticas precisas. En algunas realizaciones, se preparan y analizan de 100 a 1000 conjuntos Bootstrap. Esto está representado por unas 100 a 1000 pases a través de los bloques 729, 731 y 733 del método 725. Sin embargo, no se pretende que la presente invención se limite a un número concreto de conjuntos Bootstrap, ya que se puede usar cualquier número adecuado para el análisis deseado.

Una vez preparado y analizado un número suficiente de conjuntos de Bootstrap B, la decisión 735 se responde negativamente. Como se ha indicado, el método pasa entonces al bloque 737. Allí, se calculan la media y la desviación estándar de un coeficiente (u otro indicador generado por el modelo) para cada residuo o nucleótido (incluyendo los codones) usando los valores del coeficiente (por ejemplo, de 100 a 1000 valores, uno de cada conjunto Bootstrap). A partir de esta información, el método puede calcular el estadístico t y determinar el intervalo de confianza de que el valor medido es diferente de cero. A partir del estadístico t, calcula el valor p para el intervalo de confianza. En este caso ilustrativo, cuanto menor sea el valor p, mayor será la confianza en que el coeficiente de regresión medido es distinto de cero.

Cabe señalar que el valor p no es más que uno de los muchos tipos diferentes de caracterizaciones que pueden dar cuenta de la variación estadística de un coeficiente u otro indicador de la importancia del residuo. Los ejemplos incluyen, entre otros, el cálculo de intervalos de confianza del 95 por ciento para coeficientes de regresión y la exclusión de cualquier coeficiente de regresión para cuya consideración el intervalo de confianza del 95 por ciento cruce la línea cero. Básicamente, en algunas realizaciones, se usa cualquier caracterización que tenga en cuenta la desviación estándar, la varianza u otra medida estadísticamente relevante de la distribución de los datos. En algunas realizaciones, este paso de caracterización también tiene en cuenta la magnitud de los coeficientes.

En algunas realizaciones, da como resultado una gran desviación estándar. Esta gran desviación estándar puede deberse a varias causas, entre las que se incluyen las mediciones deficientes en el conjunto de datos y/o la representación limitada de un residuo o nucleótido concreto en el conjunto de datos original. En este último caso, algunos conjuntos Bootstrap no contendrán ninguna aparición de un residuo o nucleótido en particular. En tales casos, el valor del coeficiente para ese residuo será cero. Otros conjuntos Bootstrap contendrán por lo menos algunas apariciones del residuo o nucleótido y darán un valor distinto de cero del coeficiente correspondiente. Pero los

conjuntos que den un valor cero harán que la desviación estándar del coeficiente sea relativamente grande. Esto reduce la confianza en el valor del coeficiente y da lugar a una clasificación inferior. Pero esto es de esperar, dado que hay relativamente pocos datos sobre el residuo o nucleótido en cuestión.

A continuación, en el bloque 739, el método clasifica los coeficientes de regresión (u otros indicadores) desde el valor p más bajo (mejor) al valor p más alto (peor). Esta clasificación se correlaciona en gran medida con el valor absoluto de los propios coeficientes de regresión, debido al hecho de que cuanto mayor es el valor absoluto, más desviaciones estándar se alejan de cero. Por tanto, para una desviación típica dada, el valor p se reduce a medida que aumenta el coeficiente de regresión. Sin embargo, la clasificación absoluta no siempre será la misma tanto con el método del valor p como con el de la magnitud pura, especialmente cuando se dispone de relativamente pocos puntos de datos para empezar en el conjunto S.

Finalmente, como se muestra en el bloque 741, el método fija y alterna ciertos residuos, basándose en las clasificaciones observadas en la operación del bloque 739. Esto es esencialmente el mismo uso de clasificaciones descrito anteriormente para otras realizaciones. En un enfoque, el método fija los mejores residuos (ahora aquellos con los valores p más bajos) y alterna los otros (aquellos con los valores p más altos).

Se ha demostrado que este método 725 funciona bien in silico. Además, en algunas realizaciones, el enfoque de clasificación de valores p trata de forma natural con residuos únicos o de pocas instancias: los valores p serán generalmente más altos (peores) porque en el proceso Bootstrap, aquellos residuos que no aparecen con frecuencia en el conjunto de datos original tendrán menos probabilidades de ser recogidos aleatoriamente. Incluso si sus coeficientes son grandes, su variabilidad (medida en desviaciones estándar) será también bastante alta. En algunas realizaciones, este es el resultado deseado, ya que aquellos residuos que no están bien representados (es decir, o bien no se han visto con suficiente frecuencia o tienen coeficientes de regresión más bajos) pueden ser buenos candidatos para la alternancia en la siguiente ronda de diseño de bibliotecas.

E. GENERACIÓN DE UNA BIBLIOTECA OPTIMIZADA DE VARIANTES DE PROTEÍNAS MEDIANTE LA MODIFICACIÓN DE SECUENCIAS PREDICHAS POR MODELOS

Uno de los objetivos de la invención es generar una biblioteca de variantes de proteínas optimizada mediante evolución dirigida. Algunas realizaciones de la invención proporcionan métodos para guiar la evolución dirigida de variantes de proteínas usando los modelos secuencia-actividad generados. Los varios modelos secuencia-actividad preparados y refinados de acuerdo con los métodos descritos anteriormente son adecuados para guiar la evolución dirigida de proteínas o moléculas biológicas. Como parte del proceso, los métodos pueden identificar secuencias que se usarán para generar una nueva biblioteca de variantes de proteínas. Tales secuencias incluyen variaciones en los residuos definidos identificados anteriormente, o son precursores usados para introducir posteriormente tales variaciones. Las secuencias pueden modificarse realizando mutagénesis o un mecanismo de generación de diversidad basado en la recombinación para generar la nueva biblioteca de variantes de proteínas. La nueva biblioteca también puede usarse para desarrollar un nuevo modelo de secuencia-actividad.

En algunas realizaciones, la preparación de oligonucleótidos o secuencias de ácidos nucleicos se consigue sintetizando los oligonucleótidos o secuencias de ácidos nucleicos usando un sintetizador de ácidos nucleicos. Algunas realizaciones de la invención incluyen realizar una ronda de evolución dirigida usando los oligonucleótidos preparados o la secuencia de proteínas como bloques de construcción para la evolución dirigida. Varias realizaciones de la invención pueden aplicar recombinación y/o mutagénesis a estos bloques de construcción para generar diversidad.

Como ejemplo específico, algunas realizaciones aplican técnicas de recombinación a oligonucleótidos. En estas realizaciones, los métodos implican la selección de una o más mutaciones para una ronda de evolución dirigida evaluando los coeficientes de los términos del modelo secuencia-actividad. Las mutaciones se seleccionan a partir de combinaciones de aminoácidos o nucleótidos definidos de tipos específicos en posiciones específicas basándose en sus contribuciones a la actividad de las proteínas tal y como predicen los modelos. En algunas realizaciones, la selección de mutaciones implica la identificación de uno o más coeficientes que se determina que son mayores que otros de los coeficientes, y la selección del aminoácido o nucleótido definido en una posición definida representada por el uno o más coeficientes así identificados. En algunas realizaciones, después de seleccionar las mutaciones según los modelos de secuencia-actividad, los métodos implican preparar una pluralidad de oligonucleótidos que contengan o codifiquen la una o más mutaciones, y realizar una ronda de evolución dirigida usando los oligonucleótidos preparados. En algunas realizaciones, las técnicas de evolución dirigida implican combinar y/o recombinar los oligonucleótidos.

Otras realizaciones de la invención aplican técnicas de recombinación a secuencias de proteínas. En algunas realizaciones, los métodos implican identificar una nueva proteína o una nueva secuencia de ácido nucleico, y preparar y ensayar la nueva proteína o una proteína codificada por la nueva secuencia de ácido nucleico. En algunas realizaciones, los métodos comprenden además usar la nueva proteína o la proteína codificada por la nueva secuencia de ácido nucleico como punto de partida para una evolución dirigida posterior. En algunas realizaciones, el proceso

de evolución dirigida implica fragmentar y recombinar la secuencia de proteína que el modelo predice que tiene un nivel deseado de actividad.

En algunas realizaciones, los métodos identifican y/o preparan una nueva proteína o una nueva secuencia de ácido nucleico basándose en mutaciones individuales que el modelo predice que serán importantes. Estos métodos implican: seleccionar una o más mutaciones evaluando los coeficientes de los términos del modelo secuencia-actividad para identificar uno o más de los aminoácidos o nucleótidos definidos en las posiciones definidas que contribuyen a la actividad; identificar una nueva proteína o una nueva secuencia de ácido nucleico que comprenda la una o más mutaciones seleccionadas anteriormente, y preparar y ensayar la nueva proteína o una proteína codificada por la nueva secuencia de ácido nucleico.

En otras realizaciones, los métodos identifican y/o preparan una nueva proteína o una nueva secuencia de ácido nucleico basándose en la actividad predicha de una secuencia completa en lugar de mutaciones individuales. En algunas de estas realizaciones, los métodos implican aplicar múltiples secuencias de proteínas o múltiples secuencias de aminoácidos al modelo secuencia-actividad y determinar los valores de actividad predichos por el modelo secuencia-actividad para cada una de las múltiples secuencias de proteínas o secuencias de ácidos nucleicos. Los métodos comprenden además seleccionar una nueva secuencia de proteína o una nueva secuencia de ácido nucleico de entre las múltiples secuencias de proteínas o múltiples secuencias de aminoácidos aplicadas anteriormente evaluando los valores de actividad predichos por el modelo secuencia-actividad para las múltiples secuencias. Los métodos también comprenden preparar y ensayar una proteína que tenga la nueva secuencia de proteína o una proteína codificada por la nueva secuencia de ácido nucleico.

En algunas realizaciones, en lugar de simplemente sintetizar la única proteína mejor predicha, se genera una biblioteca combinatoria de proteínas basada en un análisis de sensibilidad de los mejores cambios en las elecciones de residuos en cada localización de la proteína. En esta realización, cuanto más sensible sea una elección de residuo dada para la proteína predicha, mayor será el cambio de idoneidad predicho. En algunas realizaciones, estas sensibilidades van de mayor a menor y las puntuaciones de sensibilidad se usan para crear bibliotecas combinatorias de proteínas en rondas posteriores (es decir, incorporando esos residuos en función de la sensibilidad). En algunas realizaciones, en las que se usa un modelo lineal, la sensibilidad se identifica simplemente considerando el tamaño del coeficiente asociado con un término de residuo dado en el modelo. Sin embargo, esto no es posible para los modelos no lineales. En cambio, en las realizaciones que utilizan modelos no lineales, la sensibilidad del residuo se determina usando el modelo para calcular los cambios en la actividad cuando se varía un solo residuo en la "mejor" secuencia predicha.

Algunas realizaciones de la invención incluyen la selección de una o más posiciones en la secuencia de proteína o secuencia de ácido nucleico y la realización de mutagénesis de saturación en la una o más posiciones identificadas de este modo. En algunas realizaciones, las posiciones se seleccionan evaluando los coeficientes de los términos del modelo secuencia-actividad para identificar uno o más de los aminoácidos o nucleótidos definidos en las posiciones definidas que contribuyen a la actividad. Por consiguiente, en algunas realizaciones, una ronda de evolución dirigida incluye la realización de mutagénesis de saturación en una secuencia de proteínas en posiciones seleccionadas usando los modelos de secuencia-actividad. En algunas realizaciones que implican modelos que comprenden uno o más términos de interacción, los métodos implican aplicar mutagénesis simultáneamente en los dos o más residuos que interactúan.

En algunas realizaciones, los residuos se toman en consideración en el orden en que se clasifican. En algunas realizaciones, para cada residuo en consideración, el proceso determina si "alternar" ese residuo. El término "alternar" se refiere a la introducción de múltiples tipos de residuos de aminoácidos en una posición específica en las secuencias de variantes de proteínas en la biblioteca optimizada. Por ejemplo, la serina puede aparecer en la posición 166 en una variante de proteína, mientras que la fenilalanina puede aparecer en la posición 166 en otra variante de proteína de la misma biblioteca. Los residuos de aminoácidos que no varían entre las secuencias de variantes de proteínas en el conjunto de entrenamiento típicamente permanecen fijos en la biblioteca optimizada. Sin embargo, no siempre es así, ya que puede haber variación en las bibliotecas optimizadas.

En algunas realizaciones, se diseña una biblioteca de variantes de proteínas optimizadas de manera que todos los residuos de coeficiente de regresión de "alta" clasificación identificados se fijan, y los restantes residuos de coeficiente de regresión de clasificación inferior se conmutan. La lógica de esta realización es que debe buscarse en el espacio local que rodea a la "mejor" proteína predicha. Cabe señalar que la "estructura principal" de punto de partida en la que se introducen los cambios puede ser la mejor proteína predicha por un modelo y/o una "mejor" proteína ya validada de una biblioteca seleccionada. De hecho, no se pretende que la estructura principal del punto de partida se limite a ninguna proteína en particular.

En una realización alternativa, se fijan por lo menos uno o más, pero no todos los residuos de coeficiente de regresión de alto rango identificados en la biblioteca optimizada, y los demás se alternan. Este enfoque se recomienda en algunas realizaciones, si existe el deseo de no cambiar drásticamente el contexto de los otros residuos de aminoácidos mediante la incorporación de demasiados cambios a la vez. De nuevo, el punto de partida para la

alternancia puede ser el mejor conjunto de residuos predicho por el modelo, una proteína mejor validada de una biblioteca existente o un clon "medio" que se modele bien. En este último caso, puede ser deseable alternar los residuos predichos como de mayor importancia, ya que se debería explorar un espacio mayor en la búsqueda de colinas de actividad previamente omitidas del muestreo. Este tipo de biblioteca es típicamente más relevante en las primeras rondas de producción de bibliotecas, ya que genera una imagen más refinada para las rondas posteriores. Tampoco se pretende que la estructura principal del punto de partida se limite a ninguna proteína en particular.

Algunas alternativas de las realizaciones anteriores incluyen diferentes procedimientos para usar la importancia de los residuos (clasificaciones) en la determinación de qué residuos alternar. En una de tales realizaciones alternativas, las posiciones de residuos mejor clasificadas se favorecen de forma más agresiva para la alternancia. La información necesaria en este enfoque incluye la secuencia de la mejor proteína del conjunto de entrenamiento, una mejor secuencia predicha por PLS o PCR, y una clasificación de residuos del modelo PLS o PCR. La "mejor" proteína es un clon validado en laboratorio húmedo como "mejor" en el conjunto de datos (es decir, el clon con la función medida más alta que todavía modela bien en el sentido de que cae relativamente cerca del valor predicho en la validación cruzada). El método compara cada residuo de esta proteína con el residuo correspondiente de una secuencia "mejor predicha" que tenga el valor más alto de la actividad deseada. Si el residuo con el coeficiente de carga o regresión más alto no está presente en el "mejor" clon, el método introduce esa posición como posición de alternancia para la biblioteca posterior. Si el residuo está presente en el mejor clon, el método no trata la posición como una posición de alternancia, y pasará a la siguiente posición en sucesión. El proceso se repite para varios residuos, moviéndose a través de valores de carga sucesivamente más bajos, hasta que se genera una biblioteca de tamaño suficiente.

En algunas realizaciones, se varía el número de residuos de coeficientes de regresión a retener y el número de residuos de coeficientes de regresión a alternar. La determinación de qué residuos alternar y cuáles retener se basa en varios factores que incluyen, pero no se limitan a, el tamaño deseado de la biblioteca, la magnitud de la diferencia entre los coeficientes de regresión y el grado en que se cree que existe no linealidad. La retención de residuos con coeficientes pequeños (neutros) puede descubrir no linealidades importantes en rondas posteriores de evolución. En algunas realizaciones, las bibliotecas de variantes de proteínas optimizadas contienen aproximadamente 2^N variantes de proteínas, donde N representa el número de posiciones que se alternan entre dos residuos. Dicho de otro modo, la diversidad añadida por cada cambio adicional duplica el tamaño de la biblioteca, de modo que 10 posiciones de cambio producen ~1.000 clones (1.024), 13 posiciones ~10.000 clones (8.192) y 20 posiciones ~1.000.000 clones (1.048.576). El tamaño adecuado de la biblioteca depende de factores como el coste de la pantalla, la robustez del paisaje, el porcentaje preferido de muestreo del espacio, etc. En algunos casos, se ha descubierto que un número relativamente grande de residuos modificados produce una biblioteca en la que un porcentaje desmesuradamente grande de los clones no son funcionales. Por lo tanto, en algunas realizaciones, el número de residuos para alternar varía entre aproximadamente 2 y aproximadamente 30; es decir, el tamaño de la biblioteca varía entre aproximadamente 4 y $2^{30} \sim 10^9$ clones.

Además, se contempla la posibilidad de utilizar simultáneamente varias estrategias de biblioteca de rondas posteriores, algunas estrategias siendo más agresivas (fijación de más residuos "beneficiosos") y otras más conservadoras (fijación de menos residuos "beneficiosos" con el objetivo de explorar el espacio más a fondo).

En algunas realizaciones, se identifican y/o conservan los grupos o residuos o "motivos" que aparecen en la mayoría de los péptidos naturales o que han tenido éxito de otra forma, ya que pueden ser importantes en la funcionalidad de la proteína (por ejemplo, actividad, estabilidad, etc.). Por ejemplo, puede encontrarse que Ile en la posición variable 3 siempre está acoplada con Val en la posición variable 11 en los péptidos naturales. Por lo tanto, en una realización, se requiere la preservación de dichos grupos en cualquier estrategia de alternancia. En otras palabras, las únicas alternancias aceptadas son aquellas que preservan una agrupación particular en la proteína base o aquellos que generan una agrupación diferente que también se encuentra en las proteínas activas. En este último caso, es necesario alternar dos o más residuos.

En algunas realizaciones adicionales, una proteína "mejor" (o una de las pocas mejores) validada en laboratorio húmedo en la biblioteca optimizada actual (es decir, una proteína con la función medida más alta, o una de las pocas más altas, que aún se modela bien, es decir, cae relativamente cerca del valor predicho en la validación cruzada) sirve como estructura principal en la que se incorporan varios cambios. En otro enfoque, una proteína "mejor" (o una de las pocas mejores) validada en laboratorio húmedo de la biblioteca actual que puede no modelizar bien sirve como estructura principal en la que se incorporan varios cambios. En algunos otros enfoques, una secuencia predicha por el modelo secuencia-actividad para tener el valor más alto (o uno de los valores más altos) de la actividad deseada sirve como estructura principal. En estos enfoques, el conjunto de datos para la biblioteca de "próxima generación" (y posiblemente un modelo correspondiente) se obtiene cambiando residuos en una o unas pocas de las mejores proteínas. En una realización, estos cambios comprenden una variación sistemática de los residuos en la estructura principal. En algunos casos, los cambios comprenden varias técnicas de mutagénesis, recombinación y/o selección de subsecuencias. Cada una de ellas puede realizarse in vitro, in vivo y/o in silico. De hecho, no se pretende que la presente invención se limite a ningún formato en particular, ya que encuentra uso cualquier formato adecuado.

En algunas realizaciones, mientras que la secuencia óptima predicha por un modelo lineal puede identificarse mediante inspección como se ha descrito anteriormente, no ocurre lo mismo con los modelos no lineales. Ciertos residuos aparecen tanto en términos lineales como de productos cruzados y su efecto global sobre la actividad en el contexto de muchas combinaciones posibles de otros residuos puede ser problemático. Por tanto, al igual que con la selección de términos de productos cruzados para un modelo no lineal, la secuencia óptima predicha por un modelo no lineal puede identificarse probando todas las secuencias posibles con el modelo (suponiendo que se disponga de suficientes recursos informáticos) o utilizando un algoritmo de búsqueda como un algoritmo por pasos.

En algunas realizaciones, la información contenida en las proteínas evolucionadas por ordenador identificadas como se ha descrito anteriormente se usa para sintetizar nuevas proteínas y probarlas en ensayos físicos. Una representación in silico precisa de la función de aptitud real determinada en el laboratorio húmedo permite a los investigadores reducir el número de ciclos de evolución y/o el número de variantes que es necesario analizar en el laboratorio. En algunas realizaciones, las bibliotecas de variantes de proteínas optimizadas se generan usando los métodos de recombinación descritos en la presente, o alternativamente, mediante métodos de síntesis génica, seguido de expresión in vivo o in vitro. En algunas realizaciones, después de que se hayan cribado las bibliotecas de variantes de proteínas optimizadas para determinar la actividad deseada, se secuencian. Como se ha indicado anteriormente en el análisis de las Figuras 1 y 2, puede emplearse la información de actividad y secuencia de la biblioteca de variantes de proteína optimizada para generar otro modelo de secuencia-actividad a partir del cual puede diseñarse otra biblioteca optimizada, usando los métodos descritos en la presente. En una realización, se usan todas las proteínas de esta nueva biblioteca como parte del conjunto de datos.

III. APARATOS Y SISTEMAS DIGITALES

Como será evidente, las realizaciones descritas en la presente emplean procesos que actúan bajo el control de instrucciones y/o datos almacenados o transferidos a través de uno o más sistemas informáticos. Las realizaciones divulgadas en la presente también se refieren a aparatos para llevar a cabo estas operaciones, como se define en las reivindicaciones. En algunas realizaciones, el aparato está especialmente diseñado y/o construido para los propósitos requeridos, o puede ser un ordenador de propósito general selectivamente activado o reconfigurado por un programa informático y/o una estructura de datos almacenada en el ordenador. Los procesos proporcionados por la presente invención no están intrínsecamente relacionados con ningún ordenador particular u otro aparato específico. En particular, varias máquinas de propósito general encuentran uso con programas escritos de acuerdo con las enseñanzas de la presente. Sin embargo, en algunas realizaciones, se construye un aparato especializado para realizar las operaciones requeridas del método. A continuación se describe una realización de una estructura particular para una variedad de estas máquinas.

Además, ciertas realizaciones de la presente invención se refieren a medios legibles por ordenador o productos de programas informáticos que incluyen instrucciones de programa y/o datos (incluyendo estructuras de datos) para realizar varias operaciones implementadas por ordenador, como se define en las reivindicaciones. Ejemplos de medios legibles por ordenador incluyen, pero no se limitan a, medios magnéticos como discos duros, disquetes, cintas magnéticas; medios ópticos como dispositivos CD-ROM y dispositivos holográficos; medios magneto-ópticos; dispositivos de memoria semiconductores; y dispositivos de hardware especialmente configurados para almacenar y ejecutar instrucciones de programa, como dispositivos de memoria de sólo lectura (ROM) y memoria de acceso aleatorio (RAM), circuitos integrados de aplicación específica (ASIC) y dispositivos lógicos programables (PLD). Los datos y las instrucciones del programa también pueden incorporarse a una onda portadora u otro medio de transporte (por ejemplo, líneas ópticas, líneas eléctricas y/o ondas aéreas). De hecho, no se pretende que la presente invención se limite a ningún medio legible por ordenador en particular ni a ningún otro producto de programa informático que incluya instrucciones y/o datos para realizar operaciones implementadas por ordenador.

Ejemplos de instrucciones de programa incluyen, pero no se limitan a código de bajo nivel como el producido por un compilador, y archivos que contienen código de nivel superior que puede ser ejecutado por el ordenador usando un intérprete. Además, las instrucciones del programa incluyen, pero no se limitan a, código máquina, código fuente y cualquier otro código que controle directa o indirectamente el funcionamiento de una máquina de computación de acuerdo con la presente invención. El código puede especificar entradas, salidas, cálculos, condicionales, ramas, bucles iterativos, etc.

En un ejemplo ilustrativo, el código que incorpora los métodos divulgados en la presente se incorpora en un medio fijo o componente de programa transmisible que contiene instrucciones lógicas y/o datos que cuando se cargan en un dispositivo informático configurado apropiadamente hace que el dispositivo realice una operación genética simulada (GO) en una o más cadenas de caracteres. La Figura 8 muestra un ejemplo de dispositivo digital 800 que es un aparato lógico que puede leer instrucciones de medios 817, puerto de red 819, teclado de entrada de usuario 809, entrada de usuario 811, u otros medios de entrada. El aparato 800 puede después usar esas instrucciones para dirigir operaciones estadísticas en el espacio de datos, por ejemplo, para construir uno o más conjuntos de datos (por ejemplo, para determinar una pluralidad de miembros representativos del espacio de datos). Un tipo de aparato lógico que puede incorporar las realizaciones divulgadas es un sistema informático como el sistema informático 800 que comprende la CPU 807, dispositivos opcionales de entrada de usuario, el teclado 809 y el dispositivo señalador de

GUI 811, así como componentes periféricos como unidades de disco 815 y el monitor 805 (que muestra cadenas de caracteres modificadas por GO y proporciona una selección simplificada de subconjuntos de dichas cadenas de caracteres por parte de un usuario. El medio fijo 817 se usa opcionalmente para programar el sistema global y puede incluir, por ejemplo, un medio óptico o magnético de tipo disco u otro elemento de almacenamiento de memoria electrónica. El puerto de comunicación 819 puede usarse para programar el sistema y puede representar cualquier tipo de conexión de comunicación.

En algunas realizaciones, como se define en las reivindicaciones, la divulgación proporciona un sistema informático que incluye uno o más procesadores; memoria del sistema; y uno o más medios de almacenamiento legibles por ordenador que tienen almacenadas en los mismos instrucciones ejecutables por ordenador que, cuando se ejecutan por el uno o más procesadores, hacen que el sistema informático implemente un método para llevar a cabo la evolución dirigida de moléculas biológicas. En algunas realizaciones, como se define en las reivindicaciones, el método incluye: (a) recibir datos de secuencia y actividad para una pluralidad de moléculas biológicas; (b) preparar un modelo base a partir de los datos de secuencia y actividad, en donde el modelo base predice la actividad como una función de la presencia o ausencia de subunidades de la secuencia; (c) preparar por lo menos un nuevo modelo añadiendo o sustrayendo por lo menos un nuevo término de interacción a o desde el modelo base, en donde el nuevo término de interacción representa la interacción entre dos o más subunidades que interactúan; (d) determinar la capacidad del por lo menos un nuevo modelo para predecir la actividad en función de la presencia o ausencia de las subunidades; y (e) determinar si añadir o sustraer el nuevo término de interacción al o del modelo base basándose en la capacidad del por lo menos un nuevo modelo para predecir la actividad según lo determinado en (d) y con un sesgo en contra de la inclusión de términos de interacción adicionales.

Ciertas realizaciones también pueden incorporarse a los circuitos de un circuito integrado de aplicación específica (ASIC) o a un dispositivo lógico programable (PLD). En tal caso, las realizaciones se implementan en un lenguaje descriptor legible por ordenador que puede usarse para crear un ASIC o PLD. Algunas realizaciones de la presente invención se implementan en los circuitos o procesadores lógicos de otros aparatos digitales, como PDA, ordenadores portátiles, pantallas, equipos de edición de imágenes, etc.

En algunas realizaciones, como se define en las reivindicaciones, la presente invención se refiere a un producto de programa de ordenador que comprende uno o más medios de almacenamiento legibles por ordenador que tienen almacenadas en los mismos instrucciones ejecutables por ordenador que, cuando se ejecutan por uno o más procesadores de un sistema informático, hacen que el sistema informático implemente un método para identificar moléculas biológicas para afectar a una actividad deseada. Dicho método puede ser cualquier método descrito en la presente, como los incluidos en las Figuras y el pseudocódigo. En algunas realizaciones, el método recibe datos de secuencia y actividad para una pluralidad de moléculas biológicas, y prepara un modelo base y un modelo mejorado a partir de los datos de secuencia y actividad. En algunas realizaciones, el modelo predice la actividad en función de la presencia o ausencia de subunidades de la secuencia.

En algunas realizaciones de la presente invención, el método implementado por el producto de programa informático prepara por lo menos un nuevo modelo añadiendo o sustrayendo por lo menos un nuevo término de interacción a o del modelo base, en donde el nuevo término de interacción representa la interacción entre dos o más subunidades que interactúan, como se define en las reivindicaciones. El método determina la capacidad de por lo menos un nuevo modelo para predecir la actividad en función de la presencia o ausencia de las subunidades. El método también determina si añadir o sustraer el nuevo término de interacción al o del modelo base basándose en la capacidad de por lo menos un nuevo modelo para predecir la actividad como se ha determinado anteriormente y con un sesgo en contra de la inclusión de términos de interacción adicionales.

REIVINDICACIONES

1. Un método implementado por ordenador de identificación de moléculas biológicas con actividad deseada mejorada, el método comprendiendo:

(a) recibir datos de secuencia y actividad para una pluralidad de moléculas biológicas;
(b) preparar un modelo base a partir de los datos de secuencia y actividad, en donde el modelo base predice la actividad en función de la presencia o ausencia de subunidades de la secuencia y el modelo base no incluye términos de interacción de un grupo definido de términos de interacción;
(b1) establecer un modelo de secuencia actual y un modelo de mejor secuencia para el modelo base;
(c1) crear un nuevo modelo de secuencia añadiendo al modelo de secuencia actual un término de interacción aún no añadido del grupo definido de términos de interacción;
(c2) evaluar la potencia predictiva del nuevo modelo de secuencia, y si la potencia predictiva del nuevo modelo de secuencia es mayor que el del mejor modelo de secuencia, establecer el nuevo modelo de secuencia como el mejor modelo de secuencia;
(c3) si hay algún término de interacción en el grupo definido de términos de interacción que no se haya añadido al modelo de secuencia actual, repetir (c1)-(c3);
(d) si se estableció un nuevo modelo como mejor modelo de secuencia en (c2), establecer el modelo de secuencia actual como mejor modelo de secuencia y repetir (c1)-(d);
(e) establecer el mejor modelo de secuencia como modelo final; y
(f) usar el modelo final para guiar la evolución dirigida y seleccionar una o más secuencias con la actividad deseada mejorada,

caracterizado por que el término de interacción se selecciona aleatoriamente en el paso (c1).

2. Un método implementado por ordenador de identificación de moléculas biológicas con actividad deseada mejorada, el método comprendiendo:

(a) recibir datos de secuencia y actividad para una pluralidad de moléculas biológicas;
(b) preparar un modelo base a partir de los datos de secuencia y actividad, en donde el modelo base predice la actividad en función de la presencia o ausencia de subunidades de la secuencia y el modelo base incluye todos los términos de interacción de un grupo definido de términos de interacción;
(b1) establecer un modelo de secuencia actual y un modelo de secuencia mejor en el modelo base;
(c1) crear un nuevo modelo de secuencia sustrayendo del modelo de secuencia actual un término de interacción aún no sustraído del grupo definido de términos de interacción;
(c2) evaluar la potencia predictiva del nuevo modelo de secuencia, y si la potencia predictiva del nuevo modelo de secuencia es mayor que la del mejor modelo de secuencia, establecer el nuevo modelo de secuencia como el mejor modelo de secuencia;
(c3) si hay algún término de interacción en el grupo definido de términos de interacción que no se haya sustraído del modelo de secuencia actual, repetir (c1)-(c3);
(d) si se estableció un nuevo modelo como mejor modelo de secuencia en (c2), establecer el modelo de secuencia actual como mejor modelo de secuencia y repetir (c1)-(d);
(e) establecer el mejor modelo de secuencia como modelo final; y
(f) usar el modelo final para guiar la evolución dirigida y seleccionar una o más secuencias con actividad deseada mejorada,

caracterizado por que el término de interacción se selecciona aleatoriamente en el paso (c1).

3. El método de la reivindicación 1 o de la reivindicación 2, en donde crear un nuevo modelo de secuencia en (c1) comprende usar información previa sobre los parámetros del nuevo modelo de secuencia para determinar las distribuciones de probabilidad posteriores de los parámetros del nuevo modelo de secuencia.

4. El método de la reivindicación 3, en donde preparar un modelo base y/o crear un nuevo modelo de secuencia comprende usar el muestreo de Gibbs para ajustar un modelo a los datos de secuencia y actividad.

5. El método de la reivindicación 1 o de la reivindicación 2, en donde la potencia predictiva del nuevo modelo de secuencia en (c2) se mide mediante el Criterio de Información de Akaike o el Criterio de Información Bayesiano.

6. El método de la reivindicación 1 o de la reivindicación 2, en donde la secuencia es un genoma completo, un cromosoma completo, un segmento cromosómico, una colección de secuencias de genes para genes que interactúan, genes o proteínas.

7. El método de cualquiera de las reivindicaciones 1 ó 2, en donde las subunidades son cromosomas, segmentos de cromosomas, haplotipos, genes, nucleótidos, codones, mutaciones, aminoácidos o residuos.

8. El método de la reivindicación 1 o de la reivindicación 2, en donde la pluralidad de moléculas biológicas constituye un conjunto de entrenamiento de una biblioteca de variantes de proteínas.

9. Un producto de programa informático que comprende uno o más medios de almacenamiento no transitorios legibles por ordenador que tienen almacenadas instrucciones en los mismos ejecutables por ordenador que, cuando son ejecutadas por uno o más procesadores de un sistema informático, hacen que el sistema informático implemente el método de cualquiera de las reivindicaciones 1-8.

10. Un sistema informático, que comprende:

uno o más procesadores;
memoria del sistema; y
uno o más medios de almacenamiento legibles por ordenador que tienen almacenadas en los mismos instrucciones ejecutables por ordenador que, cuando son ejecutadas por el uno o más procesadores, hacen que el sistema informático implemente el método de cualquiera de las reivindicaciones 1-8.

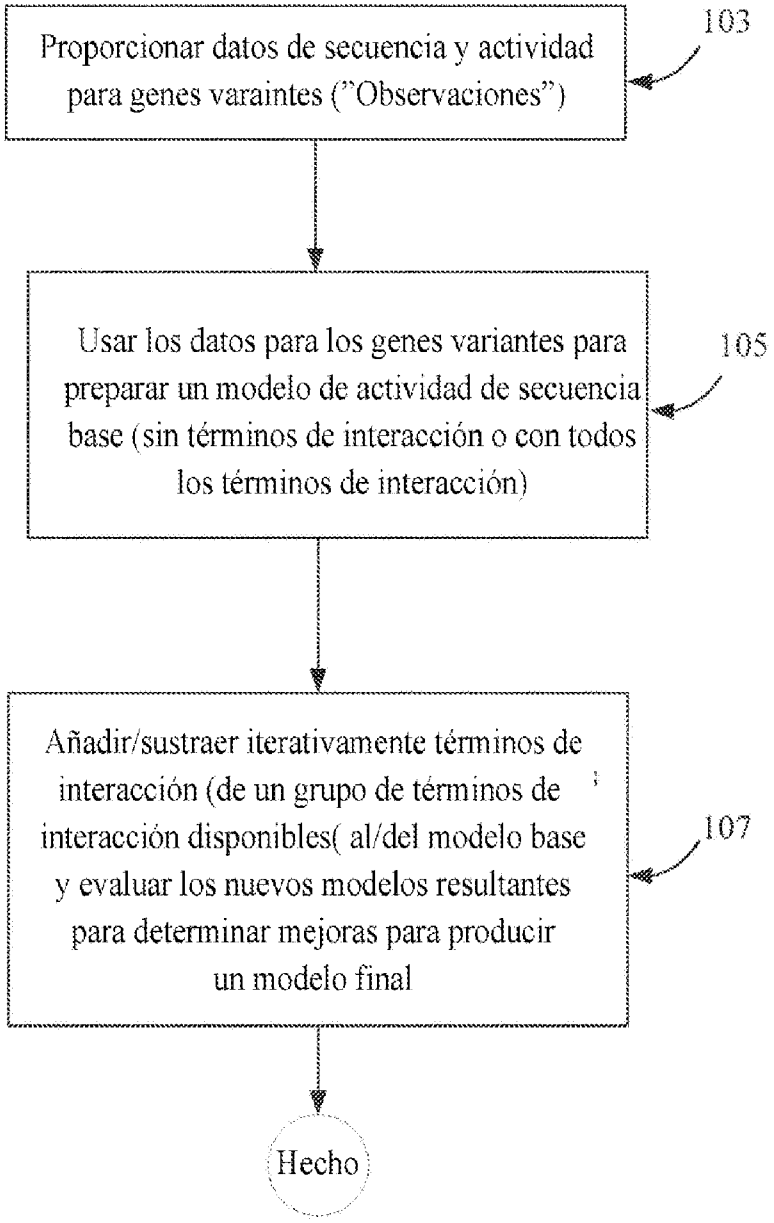
Proporcionar datos de secuencia y actividad para genes varaintes ("Observaciones")
103
Usar los datos para los genes variantes para preparar un modelo de actividad de secuencia base (sin términos de interacción o con todos los términos de interacción)
105
Añadir/sustraer iterativamente términos de interacción (de un grupo de términos de interacción disponibles(al/del modelo base y evaluar los nuevos modelos resultantes para determinar mejoras para producir un modelo final
107
Hecho

Fig. 1

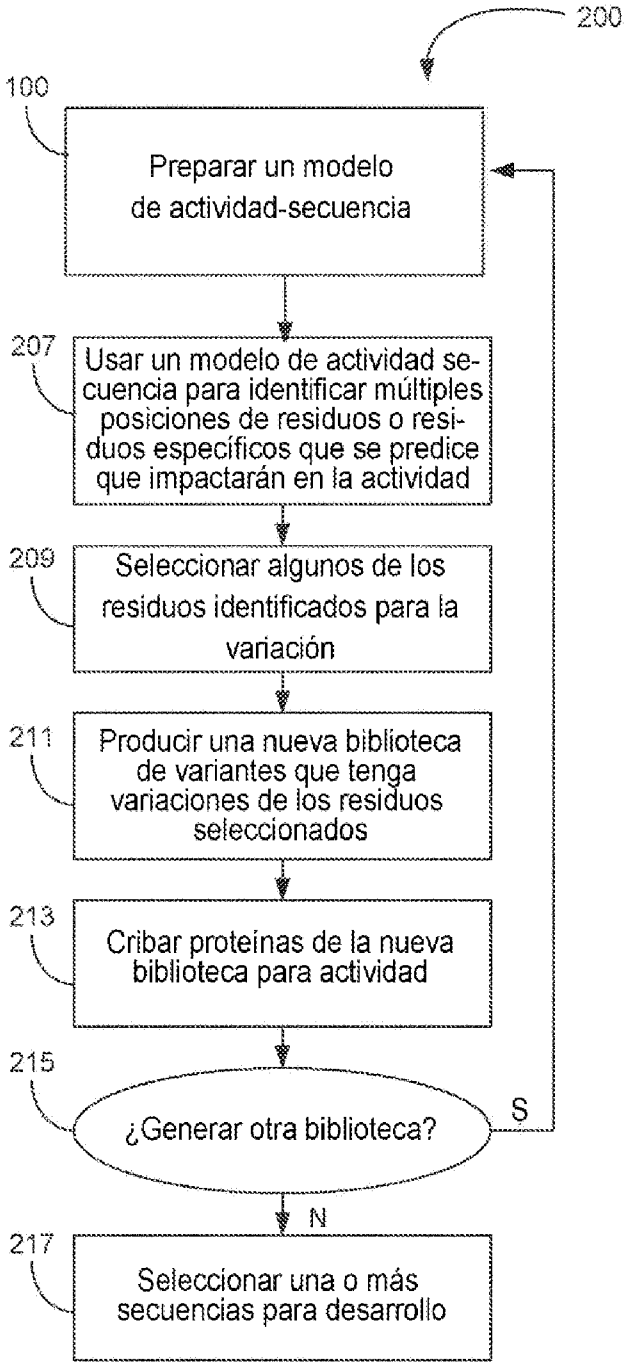
200
100
Preparar un modelo de actividad-secuencia
207
Usar un modelo de actividad secuencia para identificar múltiples posiciones de residuos o residuos específicos que se predice que impactarán en la actividad
209
Seleccionar algunos de los residuos identificados para la variación
211
Producir una nueva biblioteca de variantes que tenga variaciones de los residuos seleccionados
213
Cribar proteínas de la nueva biblioteca para actividad
215
¿Generar otra biblioteca?
S
N
217
Seleccionar una o más secuencias para desarrollo

Fig. 2

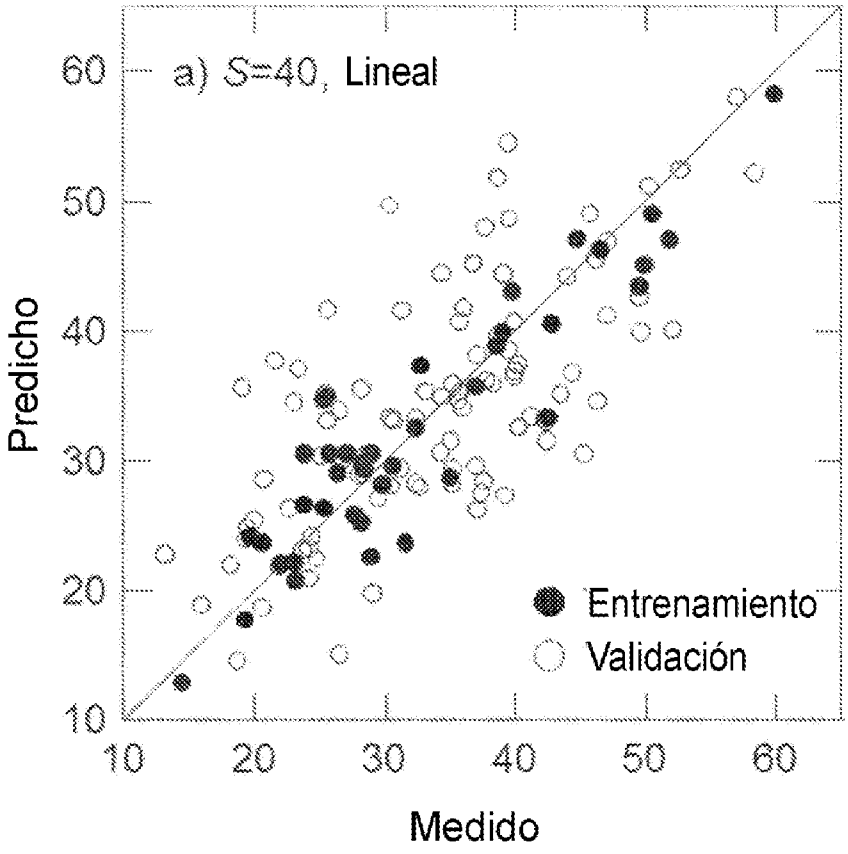
a) S=40, Lineal
Predicho
Medido
Entrenamiento
Validación
10
20
30
40
50
60

Fig. 3A

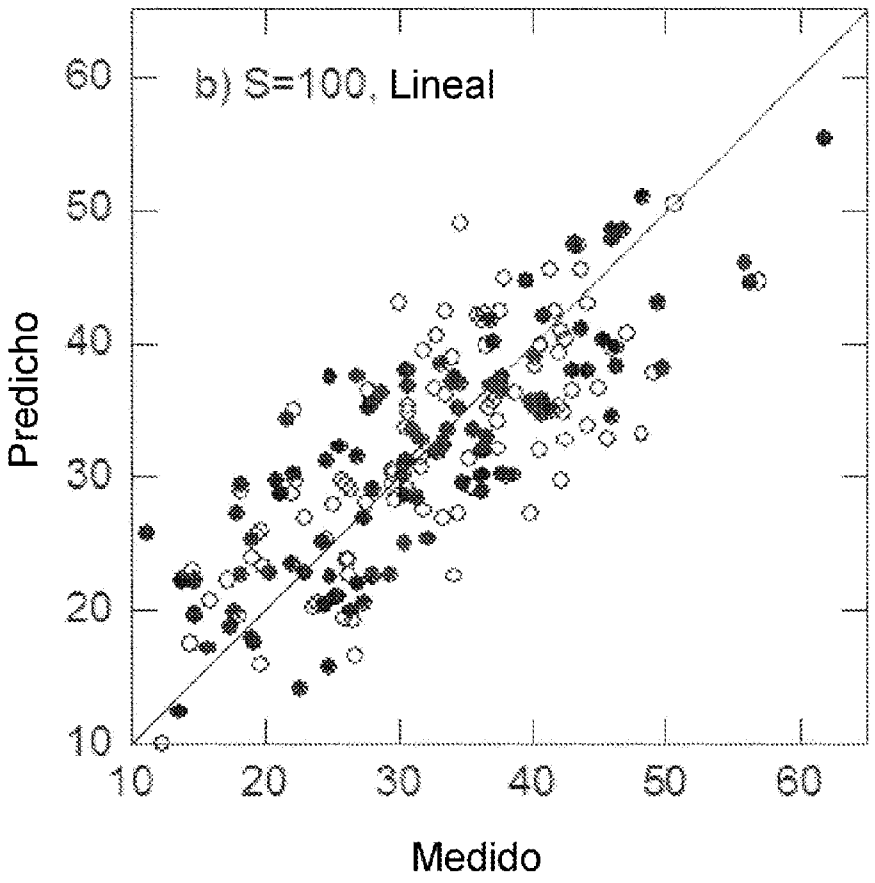

Fig. 3B

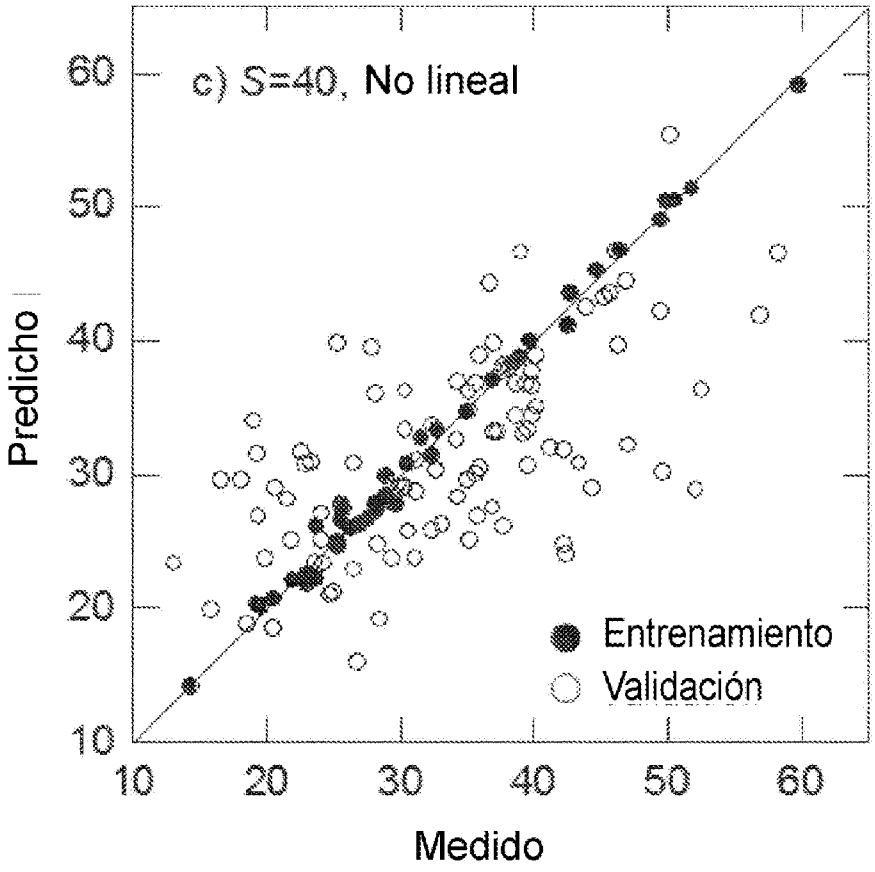
c) S=40, No lineal
Predicho
Medido
10
20
30
40
50
60
Entrenamiento
Validación

Fig. 3C

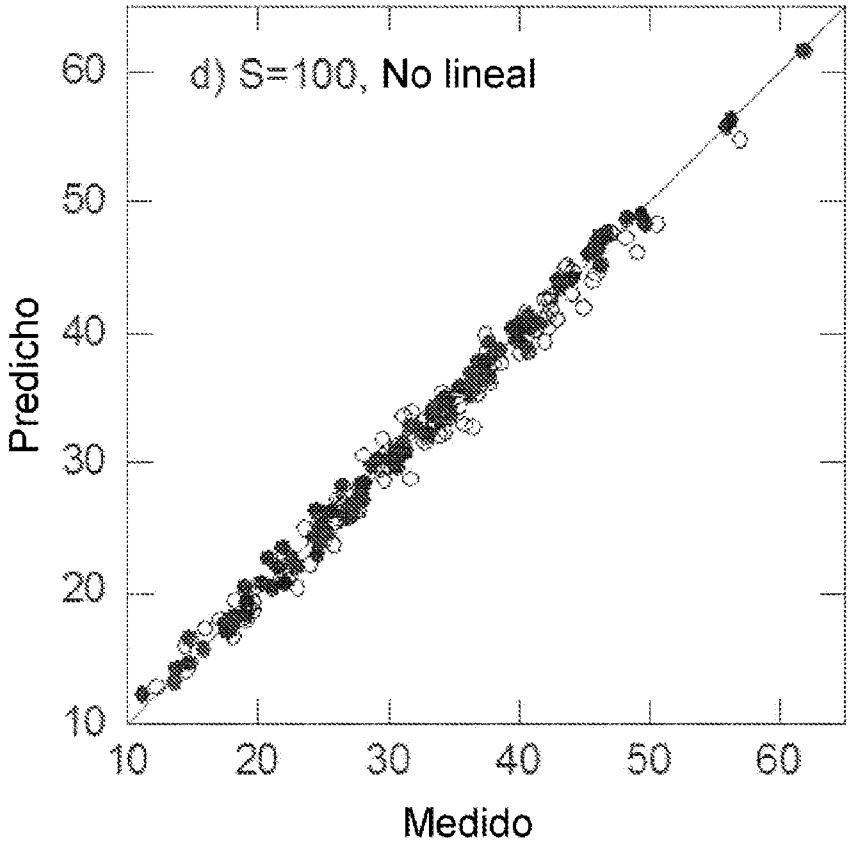

Fig. 3D

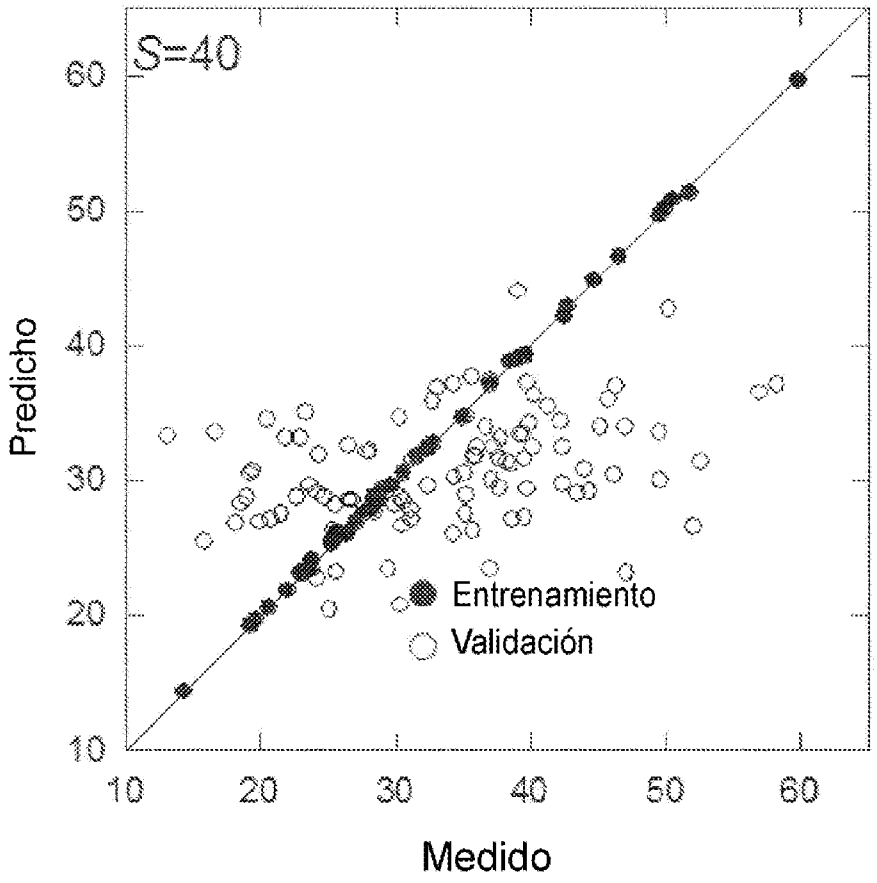
S=40
Predicho
Medido
Entrenamiento
Validación
10
20
30
40
50
60

Fig. 3E

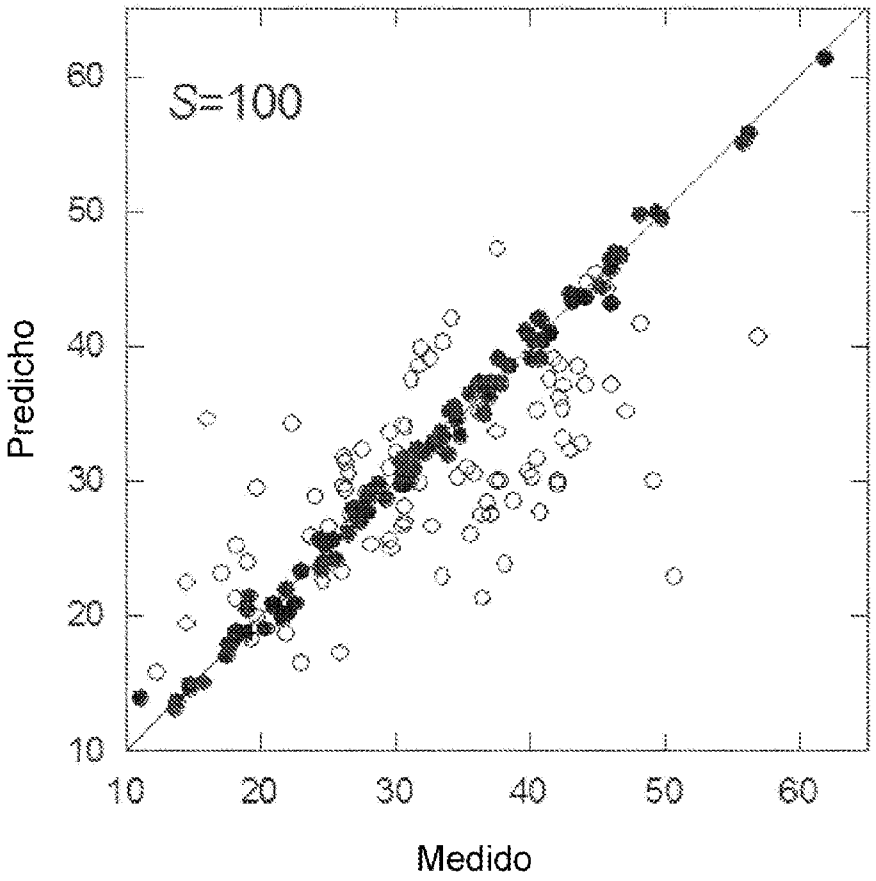

Fig. 3F

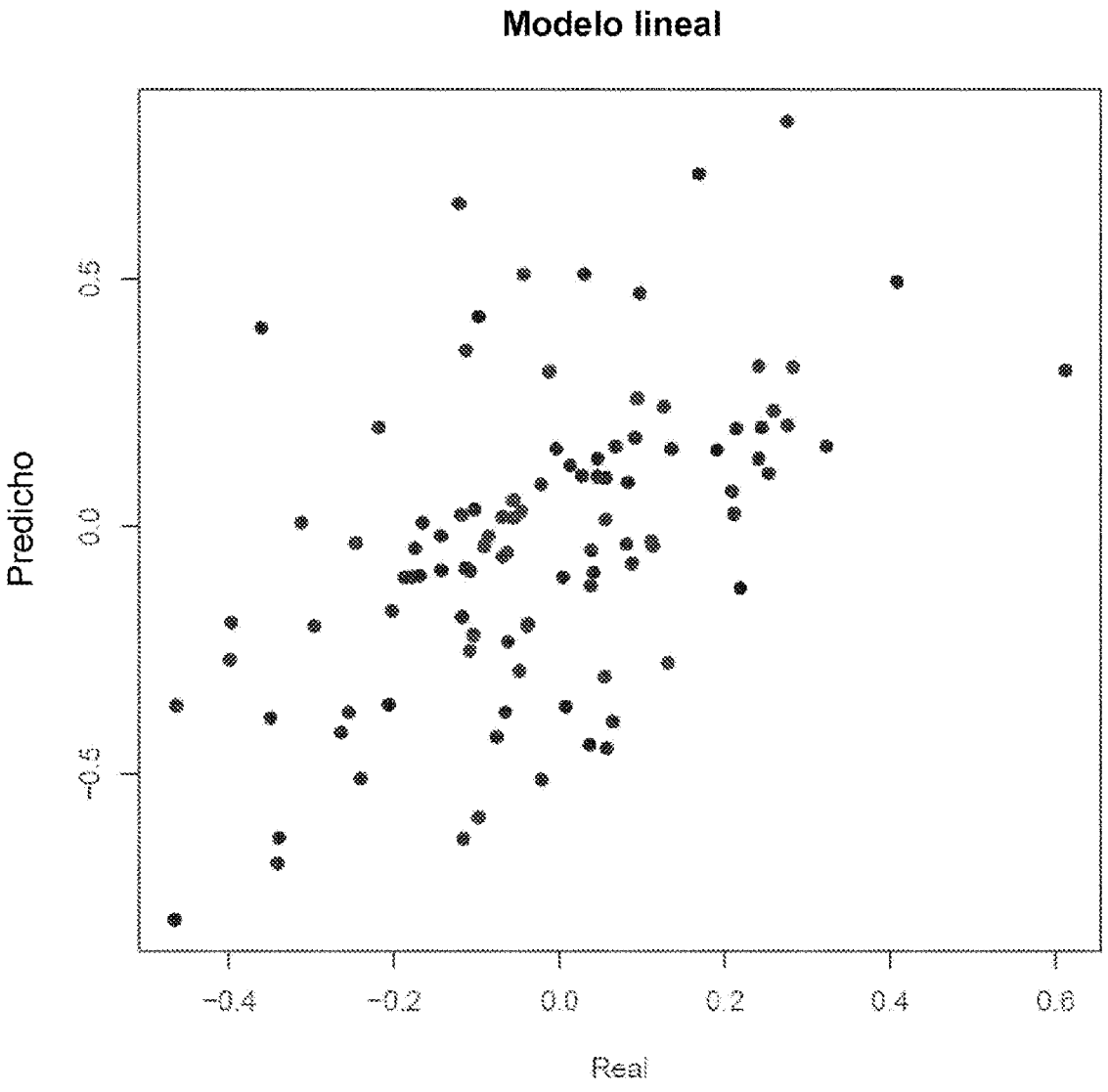

Fig. 3G

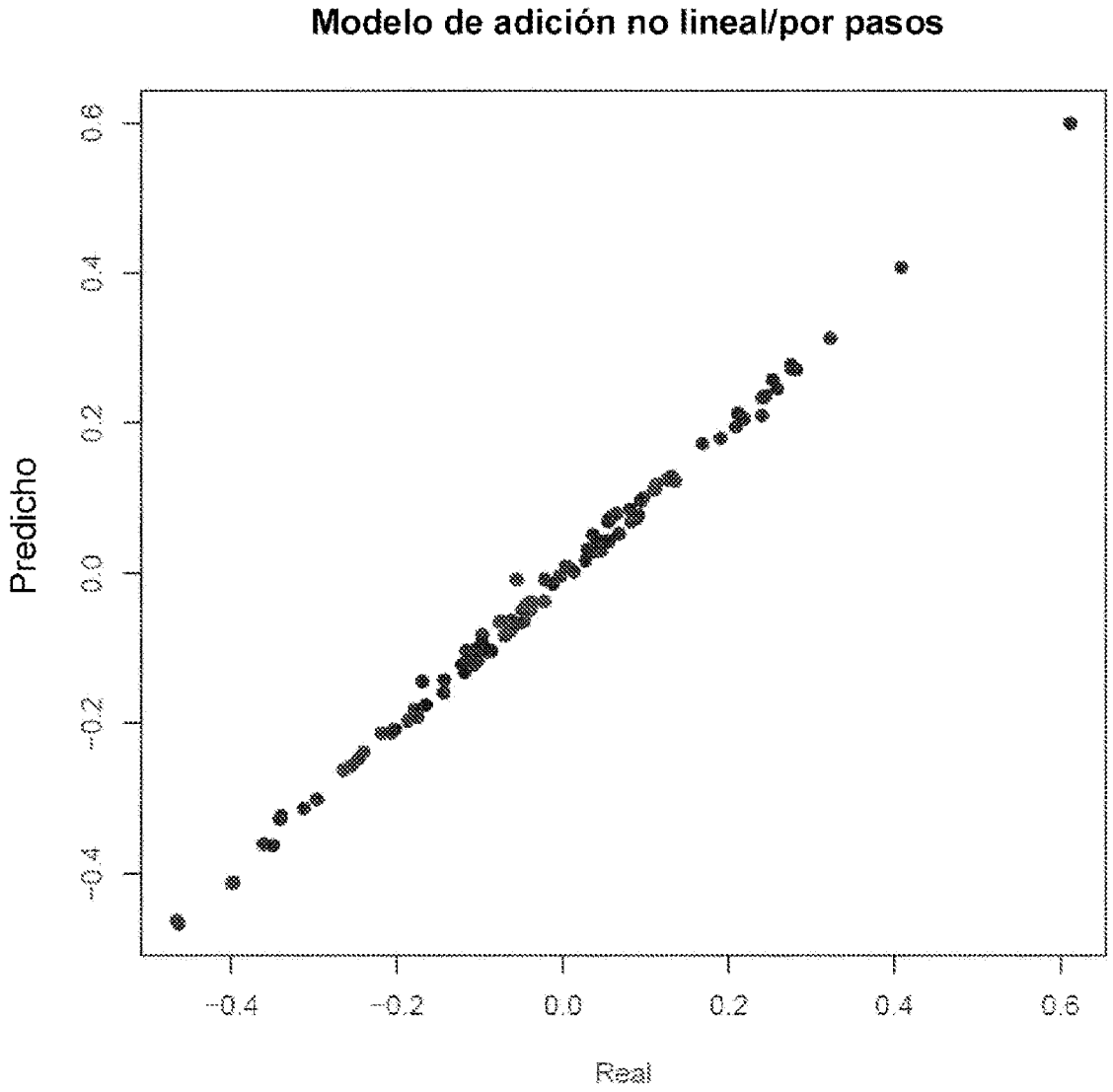
Modelo de adición no lineal/por pasos
Predicho
Real
-0.4
-0.2
0.0
0.2
0.4
0.6

Fig. 3H

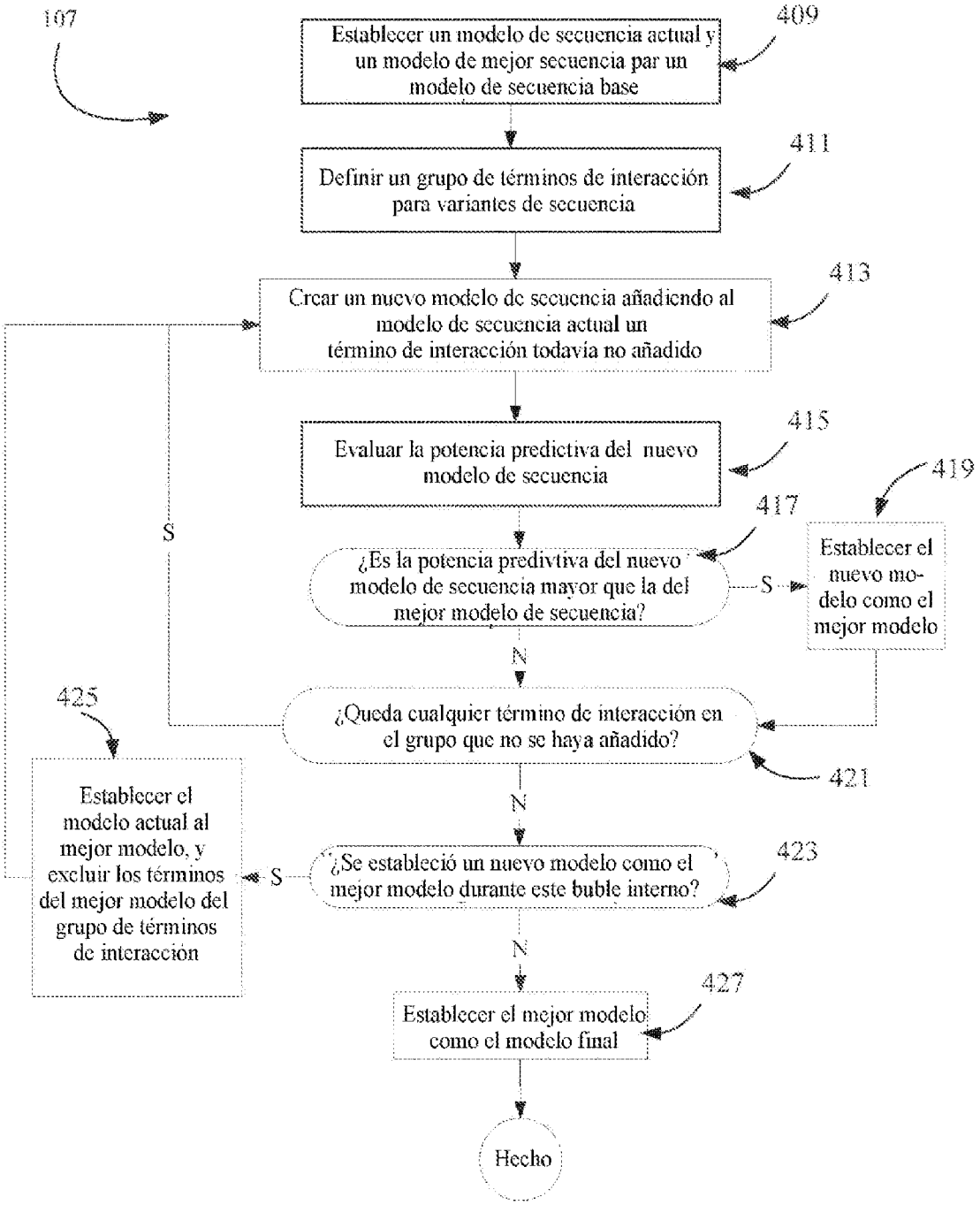
107
Establecer un modelo de secuencia actual y un modelo de mejor secuencia par un modelo de secuencia base
409
Definir un grupo de términos de interacción para variantes de secuencia
411
Crear un nuevo modelo de secuencia añadiendo al modelo de secuencia actual un término de interacción todavía no añadido
413
Evaluar la potencia predictiva del nuevo modelo de secuencia
415
¿Es la potencia predivtiva del nuevo modelo de secuencia mayor que la del mejor modelo de secuencia?
417
S
Establecer el nuevo modelo como el mejor modelo
419
N
¿Queda cualquier término de interacción en el grupo que no se haya añadido?
421
S
N
¿Se estableció un nuevo modelo como el mejor modelo durante este buble interno?
423
S
Establecer el modelo actual al mejor modelo, y excluir los términos del mejor modelo del grupo de términos de interacción
425
N
Establecer el mejor modelo como el modelo final
427
Hecho

Fig. 4A

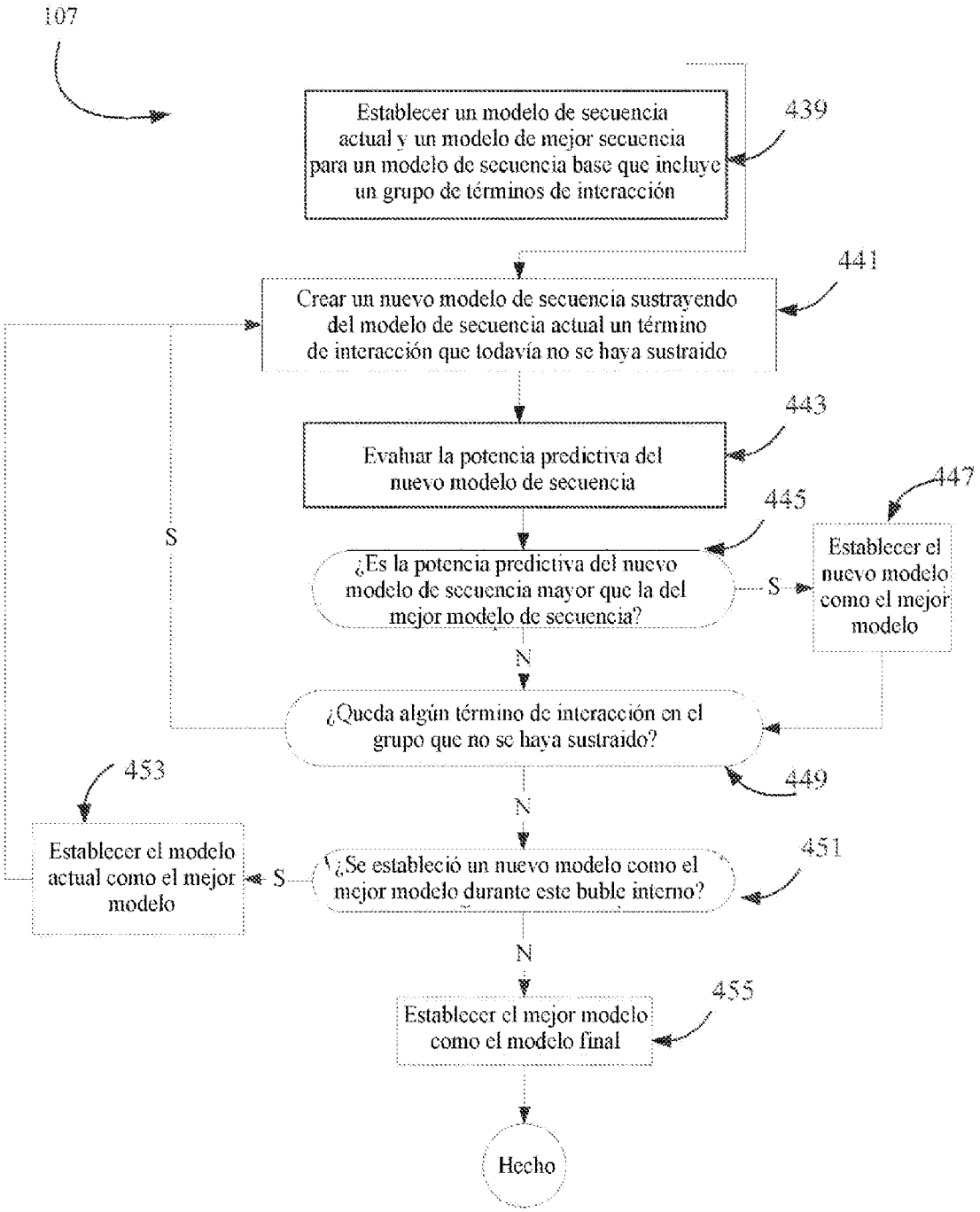
107
Establecer un modelo de secuencia actual y un modelo de mejor secuencia para un modelo de secuencia base que incluye un grupo de términos de interacción
439
Crear un nuevo modelo de secuencia sustrayendo del modelo de secuencia actual un término de interacción que todavía no se haya sustraido
441
Evaluar la potencia predictiva del nuevo modelo de secuencia
443
¿Es la potencia predictiva del nuevo modelo de secuencia mayor que la del mejor modelo de secuencia?
445
S
Establecer el nuevo modelo como el mejor modelo
447
N
¿Queda algún término de interacción en el grupo que no se haya sustraido?
449
S
N
¿Se estableció un nuevo modelo como el mejor modelo durante este buble interno?
451
S
Establecer el modelo actual como el mejor modelo
453
N
Establecer el mejor modelo como el modelo final
455
Hecho

Fig. 4B

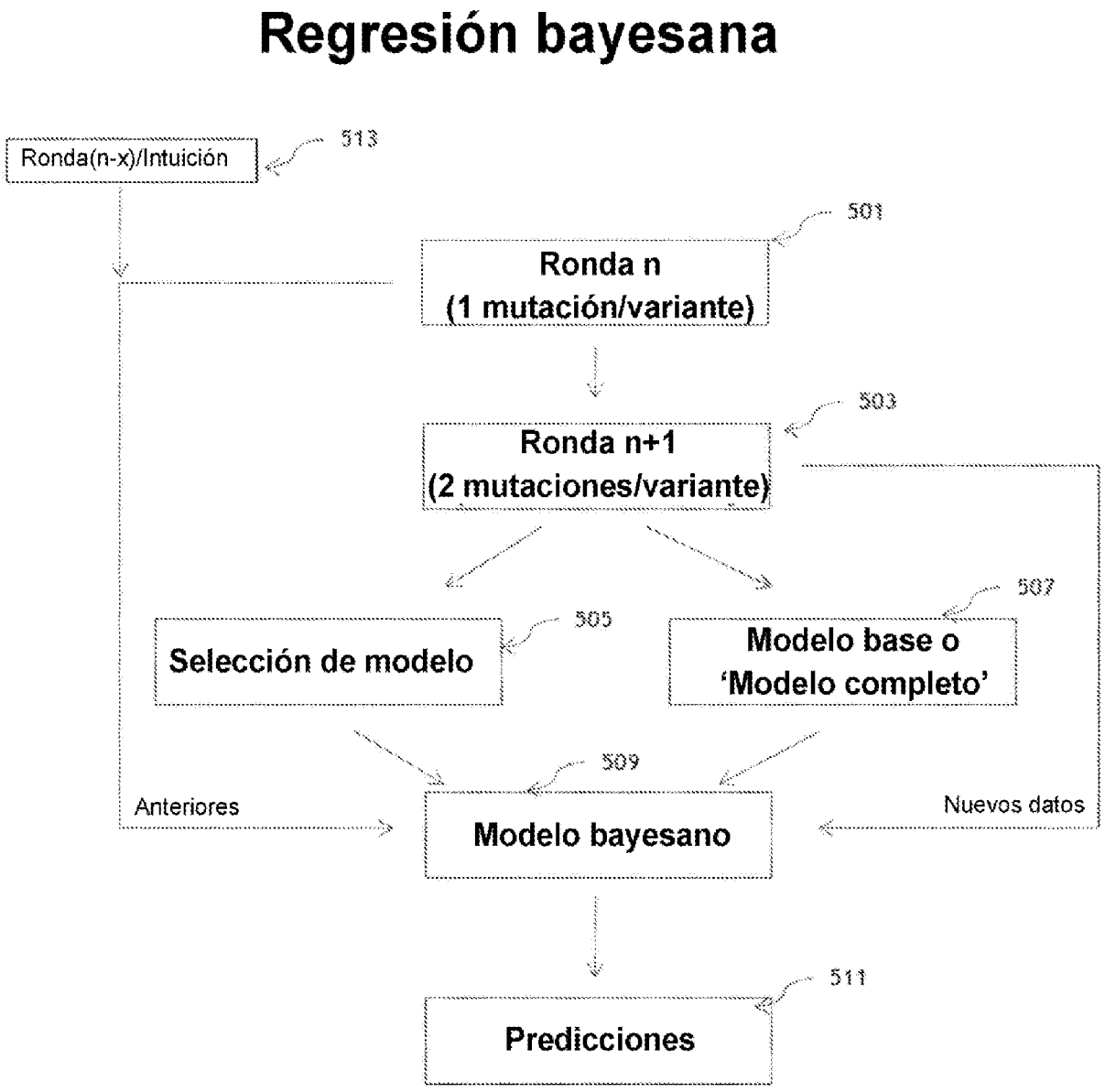
Regresión bayesana
Ronda(n-x)/Intuición
513
501
Ronda n
(1 mutación/variante)
503
Ronda n+1
(2 mutaciones/variante)
507
505
Selección de modelo
Modelo base o
'Modelo completo'
509
Anteriores
Nuevos datos
Modelo bayesano
511
Predicciones

Fig. 5

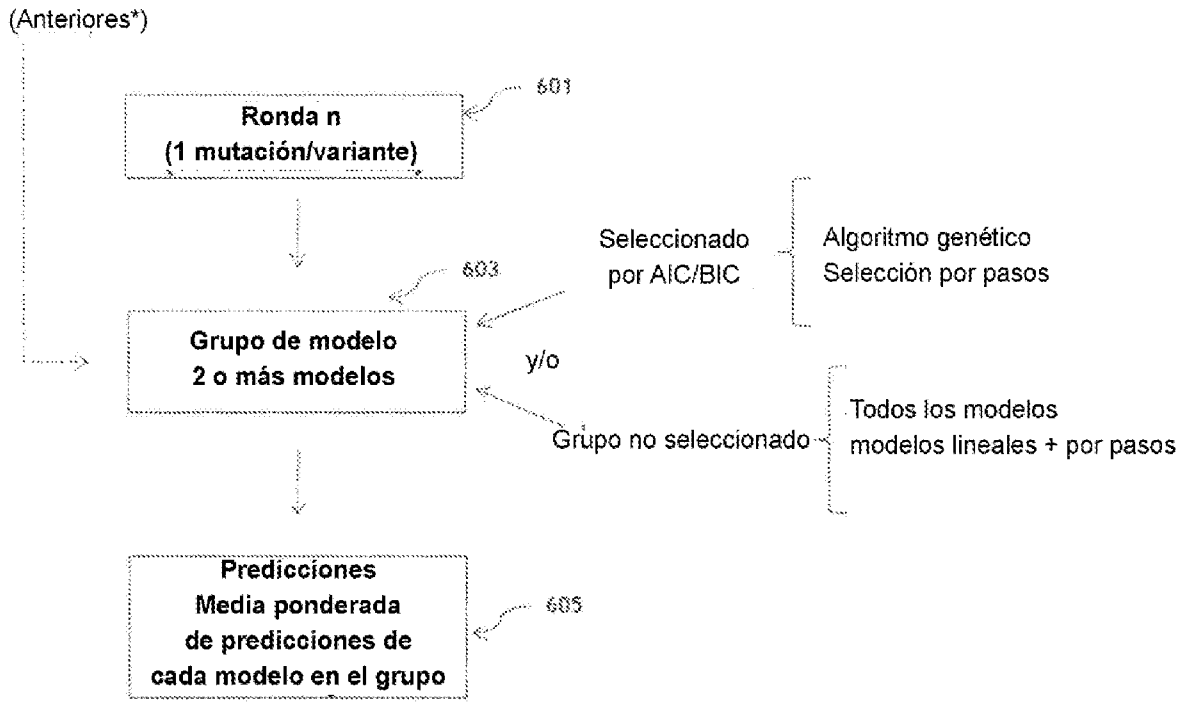
Regresión de conjunto
(Anteriores*)
601
Ronda n
(1 mutación/variante)
603
Seleccionado
por AIC/BIC
Algoritmo genético
Selección por pasos
Grupo de modelo
2 o más modelos
y/o
Grupo no seleccionado
Todos los modelos
modelos lineales + por pasos
Predicciones
Media ponderada
de predicciones de
cada modelo en el grupo
605

Fig. 6

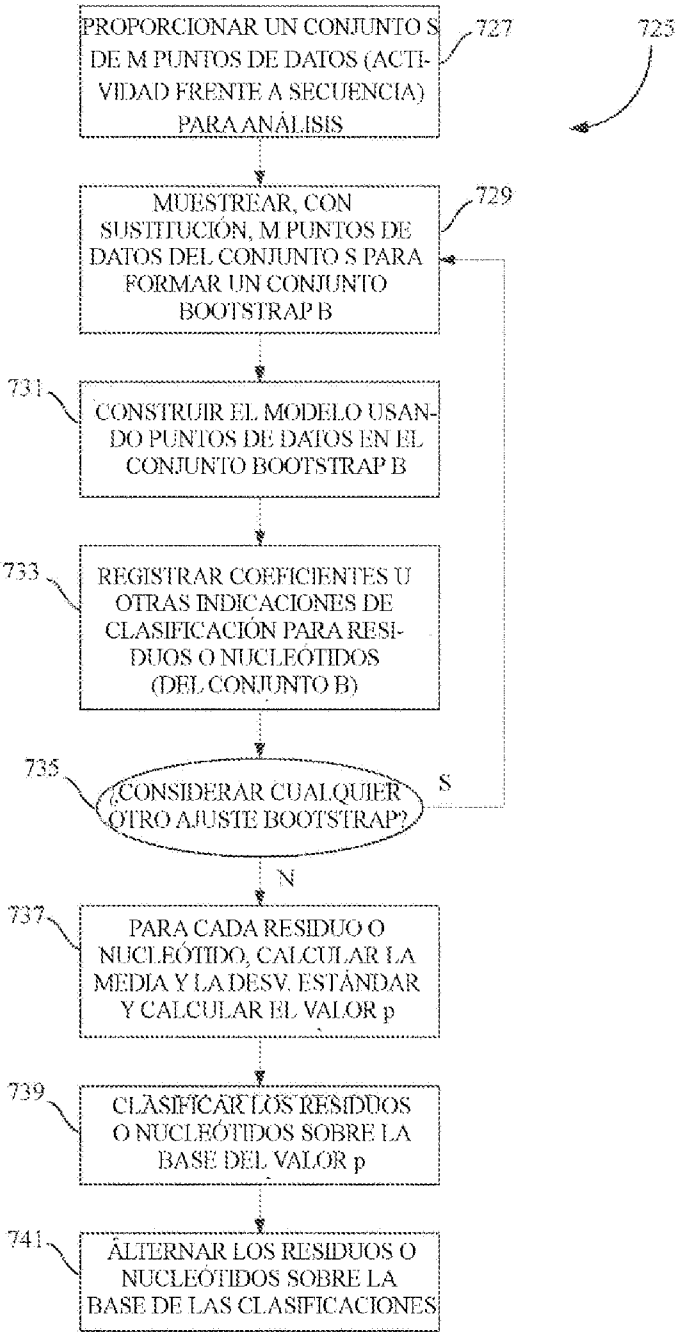
PROPORCIONAR UN CONJUNTO S DE M PUNTOS DE DATOS (ACTIVIDAD FRENTE A SECUENCIA) PARA ANÁLISIS
727
723
MUESTREAR, CON SUSTITUCIÓN, M PUNTOS DE DATOS DEL CONJUNTO S PARA FORMAR UN CONJUNTO BOOTSTRAP B
729
731
CONSTRUIR EL MODELO USANDO PUNTOS DE DATOS EN EL CONJUNTO BOOTSTRAP B
733
REGISTRAR COEFICIENTES U OTRAS INDICACIONES DE CLASIFICACIÓN PARA RESIDUOS O NUCLEÓTIDOS (DEL CONJUNTO B)
735
¿CONSIDERAR CUALQUIER OTRO AJUSTE BOOTSTRAP?
S
N
737
PARA CADA RESIDUO O NUCLEÓTIDO, CALCULAR LA MEDIA Y LA DESV. ESTÁNDAR Y CALCULAR EL VALOR p
739
CLASIFICAR LOS RESIDUOS O NUCLEÓTIDOS SOBRE LA BASE DEL VALOR p
741
ALTERNAR LOS RESIDUOS O NUCLEÓTIDOS SOBRE LA BASE DE LAS CLASIFICACIONES

Fig. 7

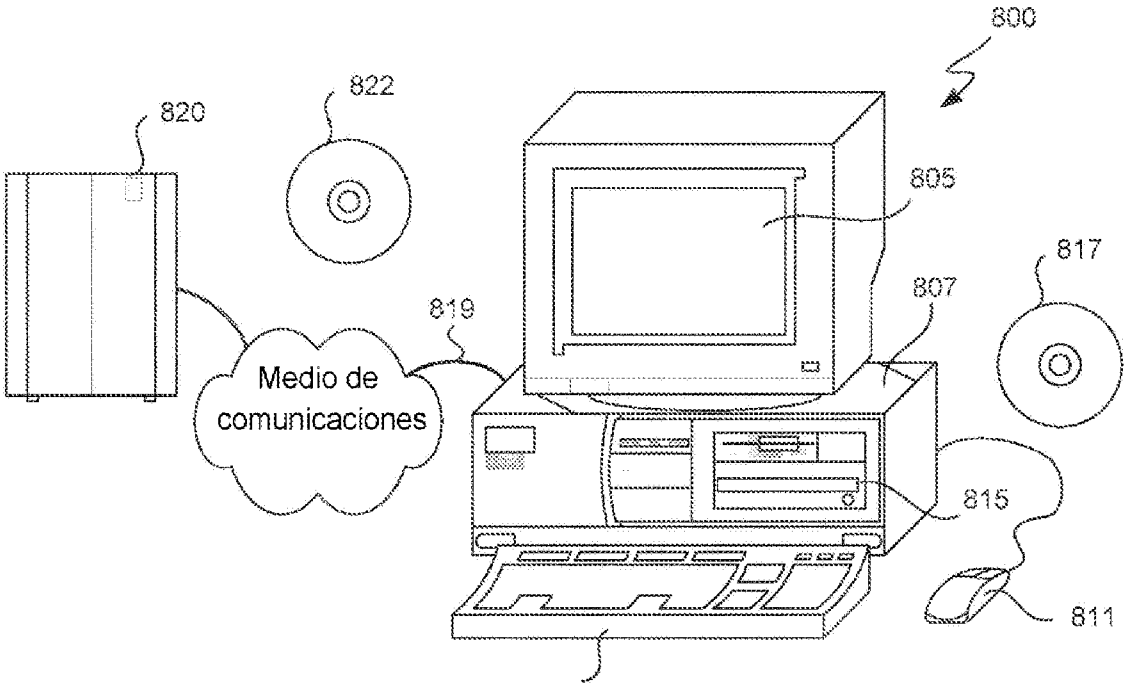

Fig. 8