(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 564 570**

(51) Int. CI.:

G06F 19/00 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **03.03.2003 E 10181159 (4)**

(97) Fecha y número de publicación de la concesión europea: **25.11.2015 EP 2315145**

(54) Título: **Métodos, sistemas y software para la identificación de biomoléculas funcionales**

(30) Prioridad:

01.03.2002 US 360982 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
23.03.2016

(73) Titular/es:

CODEXIS MAYFLOWER HOLDINGS, LLC (100.0%)
200 Penobscot Drive
Redwood City, CA 94063, US

(72) Inventor/es:

GUSTAFSSON, CLAES;
GOVINDARAJAN, SRIDHAR;
EMIG, ROBIN;
FOX, RICHARD JOHN;
ROY, AJOY;
MINSHULL, JEREMY;
DAVIS, S. CHRISTOPHER;
COX, ANTHONY;
PATTEN, PHIL;
CASTLE, LINDA A.;
SIEHL, DANIEL L.;
GORTON, REBECCA LYNNE y
CHEN, TEDDY

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Métodos, sistemas y software para la identificación de biomoléculas funcionales

CAMPO DE LA INVENCIÓN

La presente invención se refiere a los campos de la biología molecular, molecular evolución, la bioinformática, y sistemas digitales. Más específicamente, la invención se refiere a métodos de identificación de objetivos de biomoléculas con propiedades y deseados métodos para predecir computacionalmente la actividad de una biomolécula. Sistemas, incluyendo sistemas digitales y software de sistema para realizar estos métodos también se proporcionan. Los métodos de la presente invención tienen utilidad en la optimización de proteínas para uso industrial y terapéutico.

ANTECEDENTES

El diseño de proteínas durante mucho tiempo ha sido conocido por ser una tarea difícil, si no por otra razón que la explosión combinatoria de las posibles moléculas que constituyen espacio de secuencia buscable. El problema de diseño de proteínas ha demostrado recientemente que pertenecen a una clase de problemas conocidos como NP-duro (Pierce, et al. (2002) "Diseño de proteínas es NP-duro" Prot. Esp. 15(10): 779 a 782), lo que indica que no hay ningún algoritmo conocido que puede resolver tales problemas en tiempo polinómico. Debido a esta complejidad, muchos métodos aproximados se han utilizado para diseñar mejores proteínas; el principal de ellos es el método de evolución dirigida. La evolución dirigida de proteínas está hoy dominada por diversos formatos altos de selección y recombinación de rendimiento, realizados a menudo de forma iterativa (por ejemplo, Voigt et al., 2002; Dahiyat et al., 1997; WO 01/59066; y WO 01/61344).

El espacio de secuencia puede ser descrito como un espacio donde todas las proteínas posible vecinos se pueden obtener por una serie de mutaciones puntuales individuales. Smith (1970) "La selección natural y el concepto de un espacio de proteínas," Nature, 225 (232): 563-4. Por ejemplo, una proteína larga de residuos 100 sería un objeto tridimensional 100 con 20 valores posibles, es decir, los 20 aminoácidos de origen natural, en cada dimensión. Cada una de estas proteínas tiene una aptitud correspondiente en un paisaje complejo. Modelos de este tipo de "paisajes de fitness" se estudiaron por primera vez por Sewall Wright (Wright (1932) "El papel de la mutación, la consanguinidad, el cruzamiento y la selección en la evolución," Procedimientos de la Conferencia Internacional sobre Genética, 1: 356-366), pero desde entonces han sido ampliados por otros (Eigen, M. (1971) "autoorganización de la materia y la evolución de macromoléculas biológicas, "Naturwissenschaften 58 (10):465-523; Kauffman, S. et al (1987)" Hacia una teoría general de la adaptación de paseos por paisajes agrestes, "J Theor. Biol 128(1): 11-45; Kauffman, E.S., et al (1989) "El modelo NK de paisajes de fitness accidentado y su aplicación a la maduración de la respuesta inmunológica," J Theor. Biol. 141(2): 211-45; Schuster, P ., et al (1994) "Paisajes: problemas de optimización complejos y estructuras de biopolímeros", Comput Chem 18 (3):.. 295-324; Govindarajan, S. et al (1997) "La evolución de las proteínas de modelo en un paisaje de plegado. " Proteínas 29 (4):. 461-6) El espacio de secuencias de proteínas es inmenso y es imposible explorar exhaustivamente. De este modo, nuevas formas de buscar de manera eficiente el espacio de secuencias para identificar proteínas funcionales sería altamente deseable.

Deb K (algoritmos genéticos de objetivos múltiples:. Dificultades de problemas y la construcción de los problemas de prueba, 1999, Computación Evolutiva Vol 7 (3), p205-230) estudia las características de problemas que pueden causar una dificultad de algoritmo genético de objetivos múltiples (GA) para converger con el verdadero frente de Pareto-óptima.

Voigt CA et al. (Diseño Evolutivo Racional: la Teoría de la Evolución de proteínas in vitro, 2001, Advances in Protein Chemistry vol. 55 pp79-l60) analiza la evolución dirigida usando la recombinación.

Kolkman JA et al. (Directed Evolution of Protein by Exon Shuffling, 2001, Nature Biotechnology vol. 19 pp 423 a 428) discuten en formatos in vitro para el exón de barajadura y aplicarlos a la evolución dirigida de proteínas.

Ness JE et al. (Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently, 2002, Nature Biotechnology vol. 20 (12) pp1251 hasta 1255) discuten una tecnología de la ingeniería de proteínas evolutivas en el que todos los aminoácidos a partir de un conjunto de antecedentes para recombinar de manera independiente de cualquier otro aminoácido.

BREVE RESUMEN DE LA INVENCIÓN

La invención proporciona un procedimiento implementado por ordenador para predecir si las secuencias de polipéptidos hipotéticos de objetivo se tienen o pueden tener al menos una propiedad funcional deseada, comprendiendo el método: (a) identificar uno o más motivos comunes a dos o más miembros de una población de variantes de secuencia de polipéptido, usando un programa de ordenador, en el que las variantes de secuencia de polipéptido son variantes en el orden y la identidad de residuos de aminoácidos, en el que el motivo es un patrón de aminoácidos, y en el que un subconjunto de la población de variantes de secuencia de polipéptidos comprende al menos una propiedad deseada funcional y un subconjunto de la población de variantes de secuencia de polipéptido carece de la al menos una propiedad funcional deseada, para producir un conjunto de datos motivo que comprende

los motivos identificados en cada uno de los subgrupos de la población de variantes de secuencia de polipéptido; (b) correlacionar al menos un motivo a partir de los datos establecidos con motivo de al menos una propiedad funcional deseada para producir una función de motivo de puntuación, por puntuación dicha por lo menos un motivo según una frecuencia de ocurrencia en cada subconjunto de la población de variantes de secuencia de polipéptidos o de acuerdo a una frecuencia de ausencia de cada subconjunto de la población de variantes de la secuencia de polipéptido, en el que la función de puntuación de motivo es capaz de predecir si las secuencias polipeptídicas hipotéticas de objetivo se tienen o están propensos a tener al menos una propiedad funcional deseada; y (c) de secuencias de polipéptidos de puntuación de objetivo hipotéticas, utilizando la función de puntuación de motivo para predecir si las secuencias de polipéptidos de objetivo hipotéticas se tienen o pueden tener al menos una propiedad funcional deseada.

La invención también proporciona un sistema para predecir si las secuencias polipeptídicas de objetivo hipotéticas se tienen o pueden tener al menos una propiedad deseada funcional, que comprende: (a) al menos un ordenador que comprende una base de datos capaz de almacenar secuencias; y (b) el software del sistema que comprende una o más instrucciones lógicas para: (i) la identificación de uno o más motivos comunes a dos o más miembros de una población de variantes de secuencia de polipéptidos, en las variantes de secuencia de polipéptido son variantes en el orden y la identidad de aminoresiduos de ácido, en el que el motivo es un patrón de aminoácidos, y en el que un subconjunto de la población de variantes de secuencia de polipéptidos comprende al menos una propiedad funcional deseada y un subconjunto de la población de variantes de secuencia de polipéptidos carece de al menos una propiedad funcional deseada, para producir un conjunto de datos con motivos que comprende los motivos señalados en cada uno de los subgrupos de la población de variantes de secuencia de polipéptido; (ii) correlacionar al menos un motivo a partir de los datos con motivos establecidos con al menos una propiedad funcional deseada para producir una función de puntuación de motivo, por puntuación dicha por lo menos un motivo de acuerdo con una frecuencia de ocurrencia en cada subconjunto de la población de variantes de secuencia de polipéptido o de acuerdo con una frecuencia de ausencia de cada subconjunto de la población de variantes de secuencia de polipéptido, en el que la función de puntuación de motivo es capaz de predecir si las secuencias de polipeptídicas de objetivo hipotéticas se tienen o pueden tener la al menos una propiedad funcional deseada; y (iii) anotando las secuencias de polipéptidos de objetivo hipotéticas, utilizando la función de puntuación de motivo para predecir si las secuencias de polipeptídicas de objetivo hipotéticas se tienen o pueden tener al menos una propiedad funcional deseada; y (iv) opcionalmente sintetizar un polipéptido que corresponde a una secuencia de polipéptido de objetivo hipotética.

La invención también proporciona un producto de programa informático para predecir si las secuencias polipeptídicas de objetivo hipotéticas se tienen o pueden tener al menos una propiedad deseada funcional, que comprende un medio legible por ordenador que tiene una o más instrucciones lógicas para: (a) identificar uno o más motivos comunes a dos o más miembros de una población de variantes de secuencia de polipéptidos, en el que las variantes de secuencia de polipéptido son variantes en el orden y la identidad de residuos de aminoácidos, en el que el motivo es un patrón de aminoácidos, y en el que un subconjunto de la población de secuencia de polipéptidos variantes comprende la al menos una propiedad funcional deseada y un subconjunto de la población de secuencia de polipéptidos variantes carece de al menos una propiedad funcional deseada, para producir un conjunto de datos con motivos que comprende los motivos identificados en cada uno de los subconjuntos de la población de variantes de secuencia de polipéptido; (b) correlacionar al menos un motivo a partir del conjunto de datos de motivos con la propiedad al menos de una función deseada para producir una función de puntuación de motivo, por puntuación de al menos un motivo de acuerdo con una frecuencia de ocurrencia en cada subconjunto de la población de variantes de secuencia de polipéptidos o de acuerdo con una frecuencia de ausencia de cada subconjunto de la población de variantes de la secuencia de polipéptidos, en el que la función de puntuación de motivo es capaz de predecir si secuencias polipéptidas de objetivo hipotéticas se tienen o pueden tener la propiedad de al menos una función deseada; y (c) de puntuación de secuencias de polipéptidos de objetivo hipotéticos, utilizando la función de puntuación de motivo para predecir si las secuencias de polipéptidos de objetivo hipotéticos se tienen o pueden tener por lo menos una propiedad funcional deseada; y (d) sintetizar opcionalmente un polipéptido que corresponde a una secuencia de polipéptido de objetivo hipotética.

BREVE RESUMEN DE LA DIVULGACIÓN

Un aspecto de la presente descripción se refiere a métodos, aparatos, y software para la identificación de residuos de aminoácidos para la variación en una biblioteca variante de la proteína. Estos residuos se variaron a continuación, en las secuencias de variantes de proteínas en la biblioteca a fin de afectar una actividad deseada, tal como la estabilidad, la actividad catalítica, terapéutica de la actividad, la resistencia a un patógeno o toxina, toxicidad, etc. El método de este aspecto puede ser descrita por la siguiente secuencia de operaciones: (a) recibir datos que caracterizan a un conjunto de entrenamiento de una variante de proteína de biblioteca; (b) a partir de los datos, el desarrollo de un modelo de actividad de la secuencia que predice la actividad como una función del residuo de aminoácido de tipo y la posición correspondiente en la secuencia; y (c) utilizando el modelo de actividad de secuencia para identificar uno o más residuos de aminoácidos en posiciones específicas en las secuencias sistemáticamente variadas que se van a variar con el fin de impactar en la actividad deseada. En este método, las variantes de la proteína de la biblioteca pueden tener sistemáticamente variadas secuencias. Además, los datos proporciona actividad y la secuencia de información para cada variante de la proteína en el conjunto de

entrenamiento.

En algunas realizaciones, el método también incluye (d) el uso de la secuencia de modelo de actividad para identificar uno o más residuos de aminoácidos que se van a permanecer fijos (en lugar de variarse) en la nueva biblioteca de variantes de la proteína.

La biblioteca de variantes de la proteína puede incluir proteínas procedentes de diversas fuentes. En un ejemplo, los miembros incluyen proteínas naturales tales como las codificadas por los miembros de una misma familia de genes. En otro ejemplo, los miembros incluyen proteínas obtenidas mediante el uso de un mecanismo de generación de diversidad basada en la recombinación. El barajado de ADN clásico (es decir, la fragmentación mediada por recombinación de ADN) se puede realizar o barajado de ADN sintético (es decir, la recombinación mediada por oligonucleótidos sintéticos) en ácidos nucleicos que codifican todo o parte de uno o más de los antecedentes de proteínas origen natural para este propósito. En aún otro ejemplo, los miembros se obtienen mediante la realización de DOE para identificar sistemáticamente las secuencias variadas.

En general, el modelo de actividad de la secuencia puede ser de cualquier forma que prediga bien la actividad de información de la secuencia. En una realización preferida, el modelo es un modelo de regresión tal como un modelo de mínimos cuadrados parciales. En otro ejemplo, el modelo es una red neural.

Usando el modelo de actividad de secuencia para identificar los residuos para la fijación o variación puede involucrar cualquiera de las muchas diferentes técnicas de análisis posibles. En algunos casos, una "secuencia de referencia" se utiliza para definir las variaciones. Tal secuencia puede ser predicha por el modelo para tener un 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 variante de proteína original de biblioteca. A partir de la secuencia de referencia, el método puede seleccionar subsecuencias para efectuar las variaciones. Además o alternativamente, el modelo de actividad de la secuencia ocupa posiciones de los residuos (o residuos específicos en ciertas posiciones) en orden de impacto en la actividad deseada.

Uno de los objetivos del método puede ser la generación de una nueva variante de la proteína de biblioteca. Como parte de este proceso, el método puede identificar las secuencias que se van a utilizar para la generación de esta nueva biblioteca. Tales secuencias incluyen variaciones en los residuos identificados en (c) anterior, o son precursores utilizados para introducir posteriormente dichas variaciones. Las secuencias se pueden modificar mediante la realización de mutagénesis o un mecanismo de generación de diversidad basada 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 ser utilizada en el desarrollo de un nuevo modelo de actividad de la secuencia.

En algunas realizaciones, el método implica la selección de uno o más miembros de la nueva biblioteca variante de la proteína para la producción. Uno o más de éstos pueden luego sintetizarse y / o expresarse en un sistema de expresión.

Sin embargo, otro aspecto de la descripción se refiere a aparatos y productos de programas informáticos que incluyen medios legibles por máquina en la que se proporcionan instrucciones y / o arreglos de datos del programa para la implementación de los métodos y sistemas de software descritos anteriormente. Con frecuencia, las instrucciones del programa se proporcionan como código para la realización de ciertas operaciones de método. Los datos, si se emplean para implementar características de esta descripción, se puede proporcionar como estructuras de datos, tablas de bases de datos, objetos de datos, u otros arreglos apropiados de información específica. Cualquiera de los métodos o sistemas de esta descripción se puede representar, en su totalidad o en parte, ya que dichas instrucciones y / o datos de programa proporcionados en medios legibles por máquina.

Estas y otras características de la presente descripción se describirán en más detalle a continuación en la descripción detallada de la divulgación y en conjunción con las figuras siguientes.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La Figura 1 es un diagrama de flujo que representa una secuencia de operaciones, incluyendo la identificación de restos particulares para la variación, que puede utilizarse para generar una o más generaciones de bibliotecas variantes de la proteína.

La figura 2 es un gráfico que ilustra un frente convexo Pareto en una parcela de un conjunto hipotético de datos.

La Figura 3 es un gráfico que ilustra un frente de Pareto no convexo en una parcela de un conjunto hipotético de datos.

La Figura 4 es un gráfico que representa ciertas etapas llevadas a cabo en una realización de un método de

identificación de los miembros de una población de biopolímero de variantes de secuencia más adecuadas para la evolución artificial.

La Figura 5 es un gráfico que representa ciertas etapas llevadas a cabo en una realización de un método de identificación de los miembros de un conjunto de carácter biopolímero de variantes de cadena que incluyen múltiples objetivos mejorado en relación a otros miembros del conjunto de variantes de cadenas de caracteres biopolímeros.

La Figura 6 es un gráfico que representa etapas realizadas en una realización de un método de la evolución de las bibliotecas para la evolución dirigida.

La Figura 7 es un gráfico que representa ciertas etapas llevadas a cabo en una forma de realización de un método para producir una población más en forma de bibliotecas de cadenas de caracteres.

La Figura 8 es un gráfico que muestra ciertas etapas llevadas a cabo en una realización de un método de selección de posiciones de aminoácidos en una variante de polipéptido para evolucionar artificialmente.

La Figura 9 es un gráfico que muestra ciertas etapas llevadas a cabo en otra realización de un método de selección de posiciones de aminoácidos en una variante de polipéptido para evolucionar artificialmente.

La Figura 10 es un gráfico que muestra ciertas etapas llevadas a cabo en una realización de un método de identificación de los aminoácidos en los polipéptidos que son importantes para una relación polipéptido de actividad de secuencia.

La figura 11 es un gráfico que representa ciertas etapas llevadas a cabo en una realización de un método para la búsqueda de una dmanera eficiente el espacio de secuencias.

La Figura 12 es un gráfico que ilustra determinadas etapas llevadas a cabo en una realización de un método para la búsqueda de manera eficiente del espacio de secuencias.

La Figura 13 es un gráfico que muestra ciertas etapas llevadas a cabo en una realización de un método de predicción de cadenas de caracteres que incluyen propiedades deseadas.

La Figura 14 ilustra esquemáticamente un ejemplo de árbol de organización según una realización de la descripción.

La Figura 15 es un gráfico que representa ciertas etapas llevadas a cabo en una realización de un método de predicción de propiedades de cadenas de caracteres de objetivo de polipéptido.

Figura 16 es un esquema de un dispositivo digital de ejemplo.

DISCUSIÓN DETALLADA

I. DEFINICIONES

Antes de describir la presente invención en detalle, debe entenderse que esta invención no se limita a composiciones o sistemas particulares que pueden, por supuesto, varíar. Es de entenderse también que la terminología usada en este documento es con el propósito de describir realizaciones particulares solamente, y no se pretende que sean limitativos. Como se utiliza en esta memoria y reclamaciones adjuntas, las formas singulares "un", "una" y "el" incluyen referentes plurales a menos que el contenido y el contexto indique claramente lo contrario. Así, por ejemplo, la referencia a "un dispositivo" incluye una combinación de dos o más de estos dispositivos, y similares. A menos que se indique lo contrario, un "o" conjunto está destinado a ser utilizado en su sentido correcto como un operador lógico booleano, que abarca tanto la selección de funciones en la alternativa (A o B, donde la selección de A es mutuamente exclusiva de B) y la selección de las características en conjunto (A o B, donde A y B son seleccionados).

Las siguientes definiciones y las incluidos a lo largo de este suplemento de divulgación son conocidas por los expertos en la técnica.

A "bio-molécula" se refiere a una molécula que se encuentra generalmente en un organismo biológico. Moléculas biológicas preferidas incluyen macromoléculas biológicas que son típicamente de naturaleza polimérica que está compuesto de múltiples subunidades 30 (es decir, "biopolímeros"). Biomoléculas típicos incluyen, pero no se limitan a moléculas que comparten algunas características estructurales con polímeros de origen natural, tales como un Rinas (formado a partir de subunidades de nucleótidos), los ADN (formado a partir de subunidades nucleótidas), y 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 (PNAS), combinaciones de ARN y ADN (por ejemplo, chimeraplasts), o similar. Bio-moléculas también incluyen, por

ejemplo, lípidos, hidratos de carbono, u otras moléculas orgánicas que son hechas por una o más moléculas genéticamente codificables (por ejemplo, una o más enzimas o vías enzimáticas) o similares.

El término "ácido nucleico" se refiere a desoxirribonucleótidos o ribonucleótidos y polímeros (por ejemplo, oligonucleótidos, polinucleótidos, etc.) de los mismos en forma mono- o bicatenaria. A menos que se limite específicamente, el término abarca ácidos nucleicos que contienen análogos conocidos de nucleótidos naturales que tienen propiedades de unión similares al ácido nucleico de referencia y se metabolizan de una manera similar a nucleótidos de origen natural. A menos que se indique lo contrario, una secuencia de ácido nucleico particular abarca también implícitamente variantes modificadas de manera conservadora del mismo (por ejemplo, sustituciones de codones degenerados) y secuencias complementarias y, así como la secuencia indicada explícitamente. Específicamente, sustituciones de codones degenerados se pueden conseguir mediante la generación de secuencias en las que se ha seleccionado la tercera posición de uno o más (o todos) de los codones está sustituido 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 Sondas (1994) Mol Cell Probes 8: 91-98). El término ácido nucleico se usa de manera intercambiable con, por ejemplo, oligonucleótido, polinucleótido, gen, ADNc, y ARNm codificado por un gen.

Una "secuencia de ácido nucleico" se refiere a la orden y la identidad de los nucleótidos que comprenden un ácido nucleico.

Un "polinucleótido" es un polímero de nucleótidos (A, C, T, U, G, etc., o de origen natural o análogos de nucleótidos artificiales) o una cadena de caracteres que representa un polímero de nucleótidos, dependiendo del contexto. O bien el ácido nucleico dado o el ácido nucleico complementario pueden determinarse a partir de cualquier secuencia de polinucleótido especificada.

El término "gen" se utiliza ampliamente para referirse a cualquier segmento de ADN asociado con una función biológica. Por lo tanto, los genes incluyen secuencias de codificación y, opcionalmente, las secuencias reguladoras necesarias para su expresión. Los genes también incluyen segmentos de ADN opcionalmente no expresadas que, por ejemplo, forman secuencias de reconocimiento para otras proteínas. Los genes se pueden obtener de una variedad de fuentes, incluida la clonación de una fuente de interés o sintetizar a partir de información de la secuencia conocida o prevista y puede incluir secuencias diseñadas para tener parámetros deseados.

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

Los términos "polipéptido" y "proteína" se utilizan indistintamente en este documento para referirse a un polímero de residuos de aminoácidos. Típicamente, el polímero tiene al menos aproximadamente 30 residuos de aminoácidos, y usualmente al menos aproximadamente 50 residuos de aminoácidos. Más típicamente, contienen al menos aproximadamente 100 residuos de aminoácidos. Los términos se aplican a polímeros de aminoácidos en los que uno o más residuos de aminoácidos son análogos, derivados o miméticos de los correspondientes aminoácidos naturales, así como a polímeros de origen natural de aminoácidos. Por ejemplo, los polipéptidos pueden ser modificados o derivatizados, por ejemplo, por la adición de residuos de carbohidratos para formar glicoproteínas. Los términos "polipéptido" y "proteína" incluyen glicoproteínas, así como no glicoproteínas.

Un "motivo" se refiere a un patrón de subunidades en o entre las moléculas biológicas. Por ejemplo, el motivo puede referirse a un patrón de la subunidad de molécula biológica sin codificar o para un patrón de subunidad de una representación codificada de una molécula biológica.

"Revisión" se refiere al proceso en el que una o más propiedades de una o más bio-moléculas se determina. Por ejemplo, los procesos de selección típicos incluyen aquellos en los que una o más propiedades de uno o más miembros de uno o más bibliotecas se determinan.

"Selección" se refiere al proceso en el que uno o más biomoléculas son identificadas, teniendo una o más propiedades de interés. Así, por ejemplo, uno puede seleccionar una biblioteca para determinar una o más propiedades de uno o más miembros de la biblioteca. Si uno o más de los miembros de la biblioteca se identifican como poseedores de una propiedad de interés, se selecciona. La selección puede incluir el aislamiento de un miembro de la biblioteca, pero no es necesario. Además, la selección y la revisión pueden ser, y a menudo son, simultáneas.

El término "covariación" se refiere a la variación correlacionada de dos o más variables (por ejemplo, aminoácidos en un polipéptido, etc.).

"Algoritmos genéticos" son procesos que imitan procesos evolutivos. Los algoritmos genéticos (GAs) se utilizan en una amplia variedad de campos para resolver problemas que no se caracterizaron completamente o demasiado complejos para permitir la caracterización completa, pero para los cuales algún tipo de evaluación analítica está

disponible. Es decir, GAs se utilizan para resolver problemas que pueden ser evaluados por alguna medida cuantificable para el valor relativo de una solución (o al menos el valor relativo de una posible solución en comparación con otro). En el contexto de la presente invención, un algoritmo genético es un proceso para seleccionar o manipular cadenas de caracteres en un ordenador, típicamente donde la cadena de caracteres corresponde a una o más moléculas biológicas (por ejemplo, ácidos nucleicos, proteínas, PNAS, o similares).

"Evolución dirigida" o "evolución artificial" se refiere a un proceso de cambiar artificialmente una cadena de caracteres mediante la selección artificial, la recombinación, u otra manipulación, es decir, que se produce en una población reproductiva en el que hay (1) variedades de individuos, con algunas variedades (2) hereditarias, de los cuales algunas variedades (3) difieren en la aptitud (éxito reproductivo determinado por resultado de la selección para una propiedad predeterminada (característica deseada). La población reproductora puede ser, por ejemplo, una población física o una población virtual en un sistema informático.

"Los operadores genéticos" son operaciones definidas por el usuario, o conjuntos de operaciones, incluyendo cada uno un conjunto de instrucciones lógicas para la manipulación de cadenas de caracteres. Operadores genéticos se aplican a causar cambios en las poblaciones de los individuos con el fin de encontrar interesantes regiones (útil) del espacio de búsqueda (poblaciones de individuos con propiedades deseadas predeterminadas) mediante medios predeterminados de selección. Medios predeterminados (o parcialmente predeterminados) de selección incluyen herramientas computacionales (operadores que comprenden pasos lógicos guiados mediante el análisis de la información que describe las bibliotecas de cadenas de caracteres), y herramientas físicas para el análisis de las propiedades físicas de los objetos físicos, que se pueden construir (sintetizado) de la materia con el propósito de crear físicamente a una representación de la información que describe las bibliotecas de cadenas de caracteres. En una realización preferida, algunas o todas de las operaciones lógicas se realizan en un sistema digital.

Cuando se hace referencia a las operaciones en las cadenas (por ejemplo, recombinaciones, hibridaciones, elongaciones, fragmentaciones, segmentaciones, inserciones, deleciones, transformaciones, etc.), se apreciará que la operación se puede realizar en la representación codificada de una molécula biológica o en la "molécula" antes de la codificación de manera que la representación codificada captura la operación.

Una "estructura de datos" se refiere a la organización y el dispositivo opcionalmente asociado para el almacenamiento de la información, por lo general varias "piezas" de la información. La estructura de datos puede ser un simple inscripción de la información (por ejemplo, una lista) o la estructura de datos puede contener información adicional (por ejemplo, anotaciones) en relación con la información contenida en el mismo, puede establecer relaciones entre los distintos "miembros" (es decir, "piezas" de información) de la estructura de datos y pueden proporcionar consejos o enlaces a recursos externos a la estructura de datos. La estructura de datos puede ser intangible pero se convierte en tangible cuando se almacena o representados en un medio tangible (por ejemplo, papel, medio legible por ordenador, etc.). La estructura de datos puede representar diversas arquitecturas de información incluyendo, pero no limitado a las listas simples, listas enlazadas, listas indexadas, tablas de datos, índices, índices de hash, bases de datos de archivos planos, bases de datos relacionales, bases de datos locales, bases de datos distribuidas, bases de datos de cliente ligero, y de la como. En formas de realización preferidas, la estructura de datos proporciona campos suficientes para el almacenamiento de una o más cadenas de caracteres. La estructura de datos se organiza opcionalmente para permitir la alineación de las cadenas de caracteres y, opcionalmente, para almacenar información con respecto a la alineación y / o similitudes de cadena y / o diferencias de cadena. En una realización, esta información es en forma de "puntuaciones" de alineación (por ejemplo, los índices de similitud) y / o mapas de alineación que muestran la subunidad individual (por ejemplo, de nucleótidos en el caso de ácidos nucleicos) alineaciones. El término "cadena de caracteres codificada" se refiere a una representación de una molécula biológica que preserva desea secuencia / información estructural con respecto a esa molécula. Como se ha indicado, no secuencias de propiedades de biomoléculas pueden ser almacenadas en una estructura de datos y las alineaciones de tales propiedades no de secuencia, de una manera análoga a la alineación basada en secuencias se puede practicar.

Se asume generalmente que dos ácidos nucleicos tienen ascendencia común cuando demuestran similitud de secuencias. Sin embargo, el nivel exacto de similitud de secuencia para establecer la homología necesaria varía en la técnica. En general, para los propósitos de esta descripción, dos secuencias de ácido nucleico se consideran homólogas cuando compartan una identidad de secuencia suficiente para permitir que la recombinación directa ocurra entre las dos secuencias.

Una "familia filogenética" se refiere a organismos, secuencias de ácidos nucleicos, secuencias de polipéptidos, o similares que comparten una relación evolutiva común o patrón de linaje.

Una "subsecuencia" o "fragmento" es cualquier porción de una secuencia completa de ácidos o aminoácidos nucleico.

Una "biblioteca" o "población" se refiere a un conjunto de al menos dos diferentes moléculas y / o cadenas de caracteres, tales como secuencias de ácidos nucleicos (por ejemplo, genes, oligonucleótidos, etc.) o productos de expresión (por ejemplo, enzimas) de los mismos. Una biblioteca o población incluye generalmente un gran número

de moléculas diferentes. Por ejemplo, una biblioteca o población incluye típicamente al menos aproximadamente 100 diferentes moléculas, más típicamente al menos aproximadamente 1000 moléculas diferentes, y a menudo al menos unos 10.000 o más moléculas diferentes.

La "Clasificación y árboles de regresión " o "CART" se refiere a un programa de árbol de clasificación que utiliza una red de búsqueda exhaustiva de todas las posibles divisiones univariantes para encontrar las divisiones de un árbol de clasificación.

La "Varianza sistemática" se refiere a los diferentes descriptores de un elemento o conjunto de los artículos que son cambiado en diferentes combinaciones.

"Los datos sistemáticamente variados" se refieren a los datos producidos, derivado, o resultante de diferentes descriptores de un elemento o conjunto de elementos que se cambió en combinaciones diferentes. Muchos descriptores diferentes se pueden cambiar al mismo tiempo, pero en diferentes combinaciones. Por ejemplo, los datos de actividad se reunieron a partir de polipéptidos en los que se han cambiado combinaciones de aminoácidos es sistemáticamente datos variados.

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

A "hyperbox" se refiere a una región seleccionada en el espacio objetivo (por ejemplo, espacio de secuencia) que incluye al menos un individuo (por ejemplo, una bio-molécula de marcado o representación de cadena de carácter de bio-molécula) que se encuentra al menos próximo a un frente Pareto en un determinado conjunto de datos.

Los términos "secuencia" y "cadenas de caracteres" se utilizan indistintamente en el presente documento para referirse a la orden y la identidad de los residuos de aminoácidos en una proteína (es decir, una secuencia de proteínas o cadena de carácter de proteína) o a la orden e identidad de nucleótides en un ácido nucleico (es decir, una secuencia de ácido nucleico o cadena de ácido nucleico de carácter).

II. GENERACIÓN DE BIBLIOTECAS DE VARIANTES DE LA PROTEÍNA MEJORADA

De acuerdo con la presente exposición, se proporcionan varios métodos para la generación de nuevas bibliotecas de variantes de proteínas que se pueden utilizar para explorar la secuencia de proteínas y el espacio de actividad. Una característica de muchas de tales métodos es un procedimiento para identificar residuos de aminoácidos en una secuencia de proteína que se predice que afectan a una actividad deseada. Como ejemplo, tal procedimiento incluye las siguientes operaciones:

(a) recibir datos que caracterizan un conjunto de entrenamiento de unas variantes de proteína, en donde los datos proporciona información sobre la actividad y la secuencia para cada variante de la proteína en el conjunto de entrenamiento;

(b) a partir de los datos, el desarrollo de un modelo de actividad de la secuencia que predice lq actividad como una función del tipo de residuos de aminoácidos y la posición correspondiente en la secuencia;

(c) utilizando el modelo de actividad de la secuencia para identificar uno o más residuos de aminoácidos en posiciones específicas en una o más variantes de proteínas que se van a variar para impactar la actividad deseada.

Otros métodos, incluyendo ligeras variaciones de este método están dentro del alcance de la presente descripción como se expone en el presente documento.

La Figura 1 presenta un diagrama de flujo que muestra varias operaciones que se pueden realizar en el orden representado o en algún otro orden. Como se muestra, un proceso 01 comienza en un bloque 03 con la recepción de los datos que describen un conjunto de entrenamiento que comprende secuencias de residuos de una biblioteca de variante de la proteína. En otras palabras, los datos de conjunto de entrenamiento se deriva de una biblioteca variante de la proteína. Normalmente que los datos se incluyen, para cada proteína en la biblioteca, una secuencia completa o parcial residuo junto con un valor de actividad. En algunos casos, múltiples tipos de actividades (por ejemplo, constante de velocidad y estabilidad térmica) se proporcionan juntos en el conjunto de entrenamiento.

En muchas realizaciones, los miembros individuales de la biblioteca variante de la proteína representan una amplia gama de secuencias y actividades. Esto permite generar un modelo de secuencia de la actividad que tiene aplicabilidad en una amplia región de espacio de secuencia. Las técnicas para generar dichas bibliotecas diversas incluyen la variación sistemática de secuencias de proteínas y técnicas de evolución dirigida. Ambos se describen con más detalle en otra parte en este documento.

Los datos de actividad se pueden obtener mediante ensayos o pantallas apropiadamente diseñados para medir magnitudes de actividad. Dichas técnicas son bien conocidas y no son el centro de esta comunicación. Los principios para el diseño de los ensayos o las pantallas apropiadas son ampliamente conocidos. Las técnicas para la obtención de secuencias de proteínas también son bien conocidos y no son el centro de esta comunicación. La actividad se utiliza con la presente divulgación puede ser estabilidad de la proteína (por ejemplo, la estabilidad térmica). Sin embargo, muchas formas de realización importantes consideran otras actividades tales como la actividad catalítica, la resistencia a patógenos y / o toxinas, actividad terapéutica, toxicidad, y similares.

Después de que los datos de formación establecidos se han generado o adquirido, el proceso lo utiliza para generar un modelo de secuencia de actividad que predice la actividad como función de la información de la secuencia. Véase el bloque 05. Tal modelo es una expresión, algoritmo u otra herramienta que predice la actividad relativa de una proteína en particular cuando se proporciona con la información de secuencia para esa proteína. En otras palabras, la información de secuencia de proteína es una predicción de entrada y la actividad es una salida. Para muchas formas de realización de esta descripción, el modelo también puede categorizar la contribución de los diversos residuos a la actividad. Los métodos para generar tales modelos (por ejemplo, PLS) se discutirán a continuación, junto con el formato de las variables independientes (información de la secuencia), el formato de la(s) variable(s) dependientes(s) (actividad), y la forma del propio modelo (por ejemplo, , una expresión de primer orden lineal).

Un modelo generado en el bloque 05 se emplea para identificar múltiples posiciones de residuos (por ejemplo, la posición 35) o valores de los residuos específicos (por ejemplo, glutamina en la posición 35) que se predicen para impactar la actividad. Véase el bloque 07. Además de la identificación de tales posiciones, es posible que las posiciones de los residuos "rango" o valores de los residuos en función de sus contribuciones a la actividad. Por ejemplo, el modelo puede predecir que la glutamina en la posición 35 tiene el efecto más pronunciado sobre la actividad, fenilalanina en la posición 208 tiene el segundo efecto más pronunciado, y así sucesivamente. En un enfoque específico se describe a continuación, se emplean coeficientes de regresión PLS para clasificar la importancia de los residuos específicos. En otro enfoque específico, se emplea una matriz de carga PLS para clasificar la importancia de las posiciones de residuos específicos.

Después de que el proceso ha identificado residuos de que la actividad de impacto, algunos de ellos se seleccionan para la modificación, tal como indica en el bloque 09. Esto se hace con el propósito de explorar el espacio de secuencias. Los residuos son seleccionadas usando cualquiera de un número de diferentes protocolos de selección, algunas de las cuales se describirán a continuación. En un ejemplo, los residuos específicos predice que tienen el mayor impacto beneficioso sobre la actividad se conservan; en otras palabras, no se varían. Un cierto número de otros residuos predice que tienen un menor impacto son, sin embargo, para la variación seleccionada. En otro ejemplo, las posiciones de residuo que se han demostrado tener el mayor impacto en la actividad se seleccionan, pero sólo si se demuestra 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 en la actividad, pero la totalidad o la mayor parte de las proteínas con alta actividad tienen leucina en esta posición, entonces la posición 197 no se seleccionaría para la variación - en este enfoque. Todas las proteínas en una biblioteca de próxima generación tendría leucina en la posición 197. Sin embargo, si algunas "buenas" proteínas tenían valina en esta posición y otros tenían leucina, entonces el proceso elegiría variar el aminoácido en esta posición.

Una vez que se han identificado los restos de variación, el método genera a continuación una nueva biblioteca variante que tiene la variación de residuo especificado. Véase el bloque 11. Diversas metodologías están disponibles para este propósito. En un ejemplo, un mecanismo de generación de la diversidad basada en la recombinación in vivo o in vitro se lleva a cabo para generar la nueva biblioteca variante. Tales procedimientos pueden emplear oligonucleótidos que contienen secuencias o subsecuencias para la codificación de las proteínas de la biblioteca de variante parental. Algunos de los oligonucleótidos estarán estrechamente relacionados, que difieren sólo en la elección de codones para aminoácidos alternativos seleccionados para la variación en 09. El mecanismo de generación de diversidad basada en la recombinación se puede realizar para uno o múltiples ciclos. Si se utilizan múltiples ciclos, cada uno implica una etapa de selección para identificar las variantes que tienen un rendimiento aceptable para ser utilizado en un siguiente ciclo de recombinación. Esta es una forma de evolución dirigida.

En un ejemplo diferente, una secuencia de proteína de "referencia" se elige y los residuos seleccionados en 09 "se activan" para identificar los miembros individuales de la biblioteca variante. Las nuevas proteínas así identificadas son sintetizadas por una técnica apropiada para generar la nueva biblioteca. En un ejemplo, la secuencia de referencia puede ser un miembro de mejor desempeño del conjunto de entrenamiento o una "mejor" secuencia predicha por un modelo PLS 30.

En otro enfoque, el modelo de actividad de la secuencia se utiliza como una "función de aptitud" en un algoritmo genético para explorar el espacio de secuencias. Después de una o más rondas del algoritmo genético (con cada ronda usando la función de aptitud para seleccionar una o más secuencias posibles para una operación de genética), una biblioteca de prósima generación se identifica para su uso como se describe en este diagrama de flujo.

Después de que la nueva biblioteca se ha producido, se tamiza para la actividad, como se indica en un bloque 13. Idealmente, la nueva biblioteca presentará uno o más miembros con mejor actividad que se observó en la biblioteca anterior. Sin embargo, incluso sin esas ventajas, la nueva biblioteca puede proporcionar una información beneficiosa. Sus miembros pueden emplearse para la generación de modelos mejorados que dan cuenta de los efectos de las variaciones seleccionadas en 09, y por lo tanto predecir con mayor precisión la actividad a través amplias regiones del espacio de secuencias. Además, la biblioteca puede representar un pasaje en el espacio de secuencia de un máximo local hacia un máximo global (en actividad).

Dependiendo del objetivo de proceso 01, puede ser deseable generar una serie de nuevas bibliotecas de variantes de proteínas, cada uno proporcionando nuevos miembros de un conjunto de entrenamiento. El conjunto de entrenamiento actualizado se utiliza para generar un modelo mejorado. Con este fin, el proceso 01 se muestra con una operación de decisión 15, que determina si todavía otra biblioteca variante de proteína debe ser producida. Varios criterios pueden utilizarse para tomar esta decisión. Los ejemplos incluyen el número de bibliotecas de variantes de proteínas generado hasta el momento, la actividad de las principales proteínas de la biblioteca de corriente, la magnitud de la actividad deseada y el nivel de mejora observado en nuevas bibliotecas recientes.

Suponiendo que el proceso va a continuar con una nueva biblioteca, el proceso vuelve a la operación 05, donde se genera un nuevo modelo de secuencia de la actividad de la secuencia y actividad de datos obtenidos para la biblioteca variante de la proteína actual. En otras palabras, los datos de secuencia y actividad de la biblioteca variante de la proteína actual sirve como parte del conjunto de entrenamiento para el nuevo modelo (o puede servir como todo el conjunto de entrenamiento). A partir de entonces, las operaciones 07, 09, 11, 13, y 15 se llevan a cabo como se describe anteriormente, pero con el nuevo modelo.

En algún momento, en el proceso de 01, este ciclo terminará y no se generará una nueva biblioteca. En ese punto, el proceso puede simplemente terminar o una o más secuencias de una o más de las bibliotecas se pueden seleccionar para el desarrollo y / o fabricación. Véase el bloque 17.

A. ELECCIÓN DE BIBLIOTECAS DE VARIANTE DE PROTEÍNAS

Las bibliotecas de variante de la proteína son grupos de proteínas múltiples generadas por los métodos de esta descripción. Las bibliotecas de variantes de proteínas también proporcionan los datos para los conjuntos de entrenamiento utilizados para generar modelos de secuencia de actividad. El número de proteínas incluidas en una biblioteca variante de la proteína depende de la aplicación y el costo.

En un ejemplo, la biblioteca variante de la proteína se genera a partir de una o más proteínas de origen natural. En un ejemplo, estos son miembros de proteínas codificadas por una sola familia de genes. Otros puntos de partida para la biblioteca pueden ser utilizados. A partir de estas semillas o proteínas de partida, la biblioteca puede ser generada por diversas técnicas. En un caso, la biblioteca se genera por combinación aleatoria de ADN clásica (es decir, la recombinación mediada por fragmentos de ADN como se describe en Stemmer (1994) Proc. Natl. Acad. Sci. EE.UU. 10.747-10.751 y WO 95/22625) o transposición de ADN sintético (es decir, recombinación sintética mediada por oligonucleótidos como se describe en Ness et al. (2002) Nature Biotechnology 20:1251-1255 y WO 00/42561) en los ácidos nucleicos que codifican parte o la totalidad de una o más proteínas parentales. En otro caso, una única secuencia de partida se modifica de varias formas para generar la biblioteca. Preferiblemente, la biblioteca se genera variando sistemáticamente los residuos individuales. En un ejemplo, se emplea una metodología de diseño experimental (DOE) para identificar sistemáticamente las diversas secuencias. En otro ejemplo, un procedimiento de "laboratorio húmedo" tal como la recombinación mediada de oligonucleótidos se utiliza para introducir algún nivel de variación sistemática.

Tal como se utiliza aquí, el término "secuencias sistemáticamente variadas" se refiere a un conjunto de secuencias en las que se ve cada residuo en múltiples contextos. En principio, el nivel de variación sistemática se puede cuantificar por el grado en que las secuencias son ortogonales las unas de las otras (máximo diferente en comparación con la media). En la práctica, el proceso no depende de que tenga secuencias ortogonales al máximo, sin embargo, la calidad del modelo será mejorada en relación directa a la ortogonalidad de la secuencia de espacio probado. En un ejemplo sencillo, una secuencia de péptido se varió sistemáticamente mediante la identificación de dos posiciones de residuos, cada uno de los cuales puede tener uno de dos aminoácidos diferentes. Una biblioteca máximamente diversa incluye las cuatro secuencias posibles. Dicha variación sistemática máxima aumenta exponencialmente con el número de posiciones variables; por ejemplo, mediante 2N cuando hay 2 opciones en cada una de las posiciones de restos N. Los expertos en la técnica reconocerá fácilmente que la variación sistemática máxima, sin embargo, no se requiere mediante los métodos descritos allí. La variación sistemática proporciona un mecanismo para la identificación de un conjunto relativamente pequeño de secuencias para las pruebas que ofrece una buena muestra del espacio de secuencia.

Las variantes de proteínas que tienen secuencias sistemáticamente variadas se pueden obtener de una variedad de modos, usando técnicas que son bien conocidas por los expertos en la técnica. Los métodos adecuados incluyen 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 polinucleótidos que se recombinan seguido de la ligadura y / o el reensamblaje por PCR de los ácidos nucleicos. Estos métodos incluyen los descritos en, por ejemplo, Stemmer (1994) Proc Natl. Acad. Sci EE.UU. 91: 10747-10751 U.S., número de patente de EE.UU.. Nº 5.605.793, "Métodos para la recombinación in vitro", la patente de EE.UU.. Nº 5.811.238, "Métodos para la generación de polinucleótidos que tienen características deseadas por selección iterativa y recombinación," la patente de EE.UU. Nº 5.830.721, " Mutagénesis de ADN por fragmentación casual y reensamblaje," la patente de EE.UU.. Nº 5.834.252, " Reacción complementaria final de la Polimerasa," la patente de EE.UU. Nº 5.837.458, "Métodos y composiciones para ingeniería celular y metabólica", "WO/42832," recombinación de secuencias polinucleótidos, usando cebadores aleatorios o definidos, WO 98/27230 "," Métodos y composiciones para el polipéptido de ingeniería ", WO 99 / 29902, "Método de creación de secuencias de polinucleótidos y polipéptidos", y similares.

Métodos de recombinación sintéticos también son particularmente adecuados para generar bibliotecas de variantes de la proteína con la variación sistemática. En los métodos de recombinación sintéticos, una pluralidad de oligonucleótidos se sintetizan que codifican colectivamente una pluralidad de los genes a ser recombinados Típicamente, los oligonucleótidos codifican colectivamente secuencias derivadas de genes homólogos parentales. Por ejemplo, los genes homólogos de interés se alinean utilizando un programa de alineación de secuencias tal como BLAST (Atschul, et al, J. Mol. Biol. 215:403-410 (1990). Los nucleótidos correspondientes a variaciones de aminoácidos entre los homólogos se observan. Estas variaciones están opcionalmente restringidos aún más a un subconjunto del total de variaciones posibles basado en el análisis de covariación de las secuencias parentales, información funcional para las secuencias parentales, selección de cambios conservativos o no conservativos entre las secuencias parentales, u otros criterios similares. Las variaciones se aumentan opcionalmente aún más para codificar la diversidad de aminoácidos adicionales en las posiciones identificadas por, por ejemplo, el análisis de la covariación de las secuencias parentales, la información funcional para las secuencias parentales, selección de cambios conservadores o no conservadores entre las secuencias parentales, o aparente tolerancia de una posición para la variación. El resultado es una secuencia génica degenerada que codifica una secuencia de aminoácidos de consenso derivada de las secuencias de genes parentales, con nucleótidos degenerados en las posiciones que codifican variaciones de aminoácidos. Oligonucleótidos se diseñan, los cuales contienen los nucleótidos necesarios para montar la diversidad presente en el gen degenerado. Los detalles relacionados con estos enfoques se pueden encontrar en, por ejemplo, Ness et al, (2002) Nature Biotechnology 20:1251-1255, WO 00/42561, "recombinación de ácido nucleico mediada de oligonucleótidos", WO 00/42560, "Métodos para hacer cadenas de caracteres, polinucleótidos y polipéptidos que tienen las características deseadas," WO 01/75767, "selección del sitio cruzado in silico," y WO 01/64864, "Recombinación mediada de ácido nucleico monocatenario de plantilla y ácido nucleico de aislamiento de fragmentos".

Las secuencias de variantes de polinucleótidos son entonces transcritas y traducidas, ya sea in vitro o in vivo, para crear un conjunto o una biblioteca de secuencias de variantes de proteínas.

El conjunto de secuencias sistemáticamente variadas también se puede diseñar a priori utilizando métodos de diseño experimental (DOE) para definir las secuencias en el conjunto de datos. Una descripción de los métodos DOE se puede encontrar en Diamond, W. J. (2001) Diseños experimentales prácticos para ingenieros y científicos John Wiley & Sons y en "Diseño experimental práctico para ingenieros y científicos", de William J Drummond (1981) Van Nostrand Reinhold Co. Nueva York, "Estadísticas para los experimentadores" George E.P. Box, William G. Hunter y J. Stuart Hunter (1978) John Wiley and Sons, Nueva York, o, por ejemplo, en el internet en itl.nist.gov/div898/handbook/. Hay varios paquetes computacionales disponibles para llevar a cabo las matemáticas relevantes, incluyendo MatLab y experto de diseño statease. El resultado es un conjunto de datos dispersos de secuencias sistemáticamente variado y ortogonal que es adecuado para la construcción del modelo de actividad de la secuencia de la presente descripción. Conjuntos de datos basados en el DOE se pueden generar fácilmente utilizando cualquiera Plackett-Burman o diseños factoriales fraccionados. Id.

En las ciencias de ingeniería o químicas, diseños factoriales fraccionados, por ejemplo, se utilizan para definir un menor número de experimentos (que en los diseños factoriales completos) en el que un factor es varíado (conmutado) entre dos o más niveles. Las técnicas de optimización se utilizan para garantizar que los experimentos elegidos son máximamente informativos en la explicación de la variación de espacio factorial. Los mismos enfoques de diseño (por ejemplo, factorial fraccional, diseño D-óptimo) pueden ser aplicados en la ingeniería de proteínas para construir un menor número de secuencias en las que se conmuta un número dado de posiciones entre dos o más residuos. Este conjunto de secuencias sería una descripción sistemática óptima de variación presente en el espacio de secuencias de proteína en cuestión. Una vez que las actividades de las moléculas correspondientes (por ejemplo, los polinucleótidos se pueden construir a través de la síntesis de genes de acuerdo con una traducción inversa de los diseños de secuencia, a continuación, expresado como polipéptidos) se miden, un modelo PLS que tiende a ser una solución óptima, se desarrolla. Debe mencionarse que cuando no hay ninguna restricción en el número de secuencias a ser construido.

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

1) Identificar las posiciones a ser alternadas con base en los principios descritos anteriormente (presentes en las secuencias parentales, nivel de conservación, etc.)

2) Crear un experimento DOE usando uno de los paquetes estadísticos comúnmente disponibles mediante la definición del número de factores (posiciones variables), el número de niveles (opciones en cada posición), y el número de experimentos a ser realizados. El contenido de información de la matriz de salida (por lo general consta de unos y ceros que representan opciones de residuos en cada posición) depende directamente del número de experimentos a ser realizados (cuanto más mejor).
3) Utilizar la matriz de salida para construir una alineación de proteínas que codifica los unos y ceros a las opciones de residuos específicos 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(s) relevante(s).
6) Construir un modelo de los genes/proteínas probados.
7) Siga los pasos descritos anteriormente para identificar posiciones de importancia y para construir una biblioteca posterior con una mejor condición física.

Para fines demostrativos, considere una proteína funcionalmente en la que los mejores residuos de aminoácidos en las posiciones 20 se han de determinar, por ejemplo, donde hay 2 aminoácidos posibles disponibles en cada posición. En este caso, un diseño factorial de resolución IV sería apropiado. Un diseño de resolución IV se define como uno en que es capaz de elucidar los efectos de todas las variables individuales, sin efectos de dos factores la superposición de ellos. El diseño sería entonces especificar un conjunto de 40 secuencias de aminoácidos específicos que cubriría la diversidad total de 2^{20} (~1 millón) secuencias posibles. Estas secuencias se generan entonces por un protocolo estándar de síntesis de genes y la función y la aptitud de estos clones se determina.

Una alternativa a los enfoques anteriores es emplear todas las secuencias disponibles, por ejemplo, la base de datos GenBank® y otras fuentes públicas, para proporcionar la biblioteca variante de la proteína. Aunque esto conlleva una potencia de cálculo enorme, las tecnologías actuales lo hacen un enfoque factible. El mapeo de todas las secuencias disponibles proporciona una indicación de regiones espaciales de secuencia de interés.

B. LA GENERACIÓN UN MODELO DE SECUENCIA DE ACTIVIDADES Y QUE UTILIZAN ESE MODELO PARA IDENTIFICAR LAS POSICIONES DE RESIDUOS PARA LA VARIACIÓN

Como se indicó anteriormente, un modelo de secuencia de la actividad que se utiliza con la presente descripción se relaciona la información de secuencia de proteína con la actividad de la proteína. La información de secuencia de proteína utilizada por el modelo puede tomar muchas formas. Con frecuencia, se trata de una secuencia completa de los residuos de aminoácidos en una proteína; por ejemplo, HGPVFSTGGA.... En algunos casos, sin embargo, puede no ser necesario para proporcionar la secuencia de aminoácidos completa. Por ejemplo, puede ser suficiente proporcionar sólo aquellos residuos que se van a variar en un esfuerzo de investigación particular. En etapas posteriores de investigación, por ejemplo, muchos residuos pueden ser fijos y sólo regiones limitadas de espacio de secuencias aún no se han explorado. En tales situaciones, puede ser conveniente proporcionar modelos de actividad de secuencia que requieren, como entradas, sólo la identificación de los residuos en las regiones de la proteína, donde la exploración continúa. Aún más, algunos modelos pueden no requerir identidades exactas de los residuos en las posiciones de los residuos, pero en lugar de identificar una o más propiedades físicas o químicas que caracterizan el aminoácido en una posición particular de residuos. Por ejemplo, el modelo puede requerir la especificación de las posiciones de residuos por mayor, hidrofobicidad, acidez, etc. En algunos modelos, se emplean combinaciones de tales propiedades.

La forma del modelo de secuencia de la actividad puede variar ampliamente, siempre y cuando proporciona un vehículo para aproximar correctamente la actividad relativa de las proteínas en base a información de la secuencia. En general, considerará la actividad como una variable dependiente y los valores de secuencia / residuos como variables independientes. Los ejemplos de la forma matemática / lógica de los modelos incluyen expresiones lineales y no lineales matemáticos de diversos órdenes, redes neuronales, la clasificación y la regresión árboles / gráficos, la agrupación de enfoques, particionamiento recursivo, máquinas de vectores de soporte, y similares. En una realización preferida, la forma modelo es un modelo aditivo lineal en la que se suman los productos de coeficientes y valores de residuos. En otra realización preferida, el modelo de formulario es un producto no lineal de varias secuencias / términos de residuos, incluyendo ciertos residuos de productos cruzados (que representan los términos de interacción entre los residuos).

Los modelos se desarrollan a partir de un conjunto de entrenamiento de la actividad en comparación con información de la secuencia para proporcionar la relación matemática / lógica entre la actividad y la secuencia. Esta relación es típicamente validado antes de su uso para la predicción de la actividad de nuevas secuencias o residuos de importancia.

Diversas técnicas para la generación de modelos están disponibles. Con frecuencia, estas técnicas son de optimización o minimización de las técnicas. Los ejemplos específicos incluyen los mínimos cuadrados parciales, varias otras técnicas de regresión, así como técnicas de optimización de la programación genética, técnicas de redes neuronales, recursivas de partición y de vectores de soporte técnicas de la máquina. En general, la técnica debe producir un modelo que puede distinguir los residuos que tienen un impacto significativo sobre la actividad de los que no lo hacen. Preferiblemente, el modelo también debe clasificar los residuos individuales o posiciones de los

residuos en función de su impacto en la actividad.

En una realización preferida de la presente descripción, el modelo de actividad de la secuencia es un modelo de regresión por mínimos cuadrados parciales variables (PLS). PLS es un algoritmo que utiliza regresión de importancia variable X (independiente) para construir modelos predictivos basados en la multicolinealidad entre las variables y su correlación con una puntuación Y (es decir, variable dependiente). Las puntuaciones X- e Y- son seleccionadas por PLS para que la relación de pares sucesivos de puntuaciones X- y Y- sea tan fuerte como sea posible. Hand, D. J., et al. (2001) Principios de Minería de Datos (Computación Adaptiva y Aprendizaje de Máquina) Boston, MA, MIT Press. Los detalles de cómo derivar una ecuación de regresión final utilizando PLS se pueden encontrar, por ejemplo, en Geladi, et al. (1986) "regresión parcial de cuadrados mínimos: un tutorial, "Anal Chim Acta 198: 1-17.

En general, un modelo de regresión PLS empleado en la práctica de la presente divulgación tiene la siguiente forma:

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

En esta expresión, y su respuesta prevista, mientras que c_{ij} y x_{ij} son el coeficiente de regresión y el valor de bit (es decir, la elección de los residuos), respectivamente, en la posición i en la secuencia. Hay posiciones de los residuos de N en las secuencias de la biblioteca de variantes de la proteína y cada una de estas puede ser ocupada por uno o más residuos. En cualquier posición dada, puede haber j = 1 por M tipos de residuos separados. Este modelo PLS asume una relación lineal (aditivo) entre los residuos en cada posición. una versión ampliada de la ecuación 1 es:

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

Los datos de la forma de la actividad y la secuencia de información se deriva de la biblioteca inicial de variante de proteína y se utilizan para determinar los coeficientes de regresión del modelo PLS. Los valores de bit se identifican primero de una alineación de las secuencias de variantes de la proteína. Posiciones de los residuos de aminoácidos son identificadas de entre la secuencias de variantes de proteína en las que los residuos de aminoácidos en esas posiciones difieren entre las secuencias. Información de residuos de aminoácidos en algunas o todas de estas posiciones de los residuos variables se puede incorporar en el modelo de actividad de la secuencia.

La Tabla I contiene información de la secuencia en forma de posiciones de residuos variables y tipo de residuos para proteínas variantes ilustrativas 10, junto con los valores de actividad correspondientes a cada proteína variante. Entender, que estos son miembros representativos de un conjunto más grande que se requiere para generar suficientes ecuaciones para resolver todos los coeficientes. Así, por ejemplo, para las secuencias de variante de proteína ilustrativos de la Tabla I, las posiciones 10, 166, 175 y 340, son posiciones de los residuos variables y todas las demás posiciones, es decir, los que no se indica en la Tabla, contienen residuos que son idénticos entre las variantes 1-10.

Tabla 1: Secuencia ilustrativa y datos de actividad

Posiciones de variables:	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	Ile	Val	Phe	y_4
Variante 5	Ala	Ile	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}

etc.

De allí que, basándose en la ecuación 1, un modelo PLS puede ser derivado de la biblioteca sistemáticamente variada en la Tabla 1, es decir:

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

Los valores de bit (variables c) pueden ser representados ya sea como 1 o 0, reflejando la presencia o ausencia del residuo de aminoácido designado o, alternativamente, 1 o -1. Por ejemplo, el uso de la designación de 1 o 0, x_{10Ala} sería "l" para Variante 1 y "-l" para Variante 2. Los coeficientes de regresión puede así derivarse de ecuaciones basadas en la información de la secuencia de actividad para todos los variantes en biblioteca. Ejemplos de tales ecuaciones para Variantes 1-10 (usando la designación 1 o 0 para x) son los siguientes:

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

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

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

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

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

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

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

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

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

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

El conjunto completo de ecuaciones se puede resolver fácilmente utilizando PLS para determinar el valor de los coeficientes de regresión correspondientes a cada residuo y la posición de interés. En este ejemplo, la magnitud relativa del coeficiente de regresión se correlaciona con la magnitud relativa de la contribución de ese resto particular de la posición particilar a actividad. Los coeficientes de regresión pueden entonces ser clasificados u otra especificación para determinar qué residuos son más propensos a contribuir favorablemente a la actividad deseada. La Tabla II proporciona valores de los coeficientes de regresión ilustrativos correspondientes a la biblioteca sistemáticamente variada se ejemplifica en la Tabla I:

Tabla II: Orden de rango ilustrativo de coeficientes de regresión

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

La lista ordenada de rango de coeficientes de regresión se puede utilizar para construir una nueva biblioteca de variantes de proteínas que se ha optimizado con respecto a una actividad deseada (es decir, estado físico mejorado). Esto se puede hacer de varias maneras. En un caso, se lleva a cabo mediante la retención de los residuos de aminoácidos que tienen coeficientes con los valores observados más altos. Estos son los residuos que el modelo PLS indica contribuyen más a la actividad deseada. Si se emplean descriptores negativos para identificar los residuos (por ejemplo, 1 para 5 leucina y -1 para la glicina), se hace necesario clasificar las posiciones de residuos basándose en el valor absoluto del coeficiente. Tenga en cuenta que en este tipo de situaciones, normalmente sólo hay un único coeficiente para cada residuo. El valor absoluto de la magnitud coeficiente da el ranking de la posición del residuo correspondiente. Entonces, se hace necesario tener en cuenta los signos de los residuos individuales para determinar si cada uno de ellos es perjudicial o beneficioso en términos de la actividad deseada.

Los residuos se consideran generalmente en el orden en el que se clasifican. Para cada residuo de que se trate, el proceso determina si se debe "cambiar" 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 de la biblioteca optimizada. Por ejemplo, la serina puede aparecer en la posición 166 en una variante de la proteína, mientras que la fenilalanina puede aparecer en la posición 166 en otra variante de la proteína en la misma biblioteca. Los residuos de aminoácidos que no varían entre las secuencias de variante de la proteína en el conjunto de entrenamiento típicamente permanecen fijos en la biblioteca optimizada.

Una biblioteca variante de la proteína optimizado puede ser diseñada de tal manera que todos los residuos de coeficientes de regresión de "alto" ranking identificados se fijan, y los residuos de los coeficientes restantes de regresión de rango inferior se alternan. La razón de esto es que uno debe buscar el espacio local que rodea a la "mejor" proteína predicha. Tenga en cuenta que el punto "columna vertebral" de partida en el que se introducen las alternaciones puede ser la mejor proteína PLS predicha o una "mejor" proteína ya validada a partir de una biblioteca filtrada.

En un enfoque alternativo, por lo menos uno o más, pero no todos los residuos de alto rango de coeficientes de regresión identificados pueden ser fijados en la biblioteca optimizada, y los otros alternados. Este método se recomienda si se desea 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 conmutación puede ser el mejor conjunto de residuos como se predijo por el modelo de PLS o una proteína mejor validado de una biblioteca existente. O el punto de partida puede ser un clon de "promedio" que modela bien. En este caso, puede ser deseable alternar los residuos que se prevea puedan ser de mayor importancia. La razón de esto es que uno debe explorar un espacio más grande en la búsqueda de colinas de actividad omitidas previamente a partir de la toma de muestras. Este tipo de biblioteca es típicamente más relevante en las primeras rondas, ya que genera una imagen más refinada para las rondas posteriores.

El número de residuos de los coeficientes de regresión de alto valor para retener, y el número de residuos de bajo valor del coeficiente de regresión para alternar, se puede variar. Factores a considerar incluyen el tamaño de la biblioteca deseada y la magnitud de la diferencia entre los coeficientes de regresión. Bibliotecas de variantes de proteínas optimizadas típicas contienen variantes de la proteína 2N, donde N representa el número de posiciones que se cambia entre dos residuos. Dicho de otra manera, la diversidad añadida por cada palanca adicional duplica el tamaño de la biblioteca de forma que las posiciones de palanca 10 produce ~ 1.000 clones (1024), 13 posiciones ~ 10.000 clones (8.192) y 20 posiciones -1,000,000 clones (1.048.576). El tamaño apropiado de la biblioteca depende de factores tales como el costo de la pantalla, la aspereza del paisaje, el muestreo de porcentaje preferido del espacio, etc.

En la práctica, se puede poner en marcha diferentes estrategias de biblioteca redonda posteriores al mismo tiempo, con algunas estrategias siendo más agresivas (fijándose más los residuos "beneficiosos") y otras estrategias siendo más conservadoras (la fijación de un menor número de residuos "beneficiosas" con la esperanza de explorar el espacio más a fondo).

Bibliotecas variantes de la proteína optimizadas pueden generarse utilizando los métodos de recombinación descritos en este documento, o alternativamente, por métodos de síntesis de genes, seguidos por expresión in vivo o in vitro. Las bibliotecas de variantes de proteínas optimizadas luego se investigan para la actividad deseada, y se secuencian. Como se ha indicado anteriormente en la discusión de la Figura 1, la información de la actividad y la secuencia de la biblioteca variante de proteína optimizada se puede emplear para generar otro modelo de actividad de la secuencia de la que una biblioteca optimizada puede ser diseñada, utilizando los métodos descritos en el presente documento. En un enfoque, todas las proteínas de esta nueva biblioteca se utilizan como parte del conjunto de datos.

En diversos enfoques, una "mejor" (o una de las pocas mejores) proteína validada en laboratorio húmedo de en la biblioteca optimizada (es decir, una proteína de la más alta, o uno de los pocos de función más elevada, que aún modela bien, es decir, llega relativamente cerca del valor predicho de la validación cruzada PLS) pueden servir como columna vertebral donde se incorporan varios esquemas de cambios. En este enfoque, el conjunto de datos para la biblioteca "próxima generación" (y posiblemente un modelo PLS correspondiente) se obtiene cambiando de residuos en una o unas pocas de las mejores proteínas de la biblioteca optimizada actual. En una realización, estos cambios comprenden una variación sistemática de los residuos en la cadena principal.

Múltiples otras variaciones en el enfoque anterior están dentro del alcance de esta descripción. Como un ejemplo, las variables xij 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 ....). Ejemplos de tales propiedades incluyen lipofilia, a granel, y las propiedades electrónicas (por ejemplo, carga formal, superficie asociada a una carga parcial van der Waals, etc.). Para aplicar este enfoque, los valores XJJ que representan los residuos de aminoácidos pueden presentarse en términos de sus propiedades o componentes principales construidos a partir de las propiedades.

Otras variaciones del enfoque anterior implican el uso de diferentes técnicas de residuos de clasificación o de lo contrario los caracterizan en términos de importancia. En el enfoque anterior, se utilizaron las magnitudes de los coeficientes de regresión para clasificar residuos. Residuos que tienen coeficientes con grandes magnitudes (por ejemplo, 166 Ile) eran vistos como los residuos de alto rango. Esta caracterización se utiliza para decidir si o no para variar un residuo en particular en la generación de una nueva biblioteca optimizada de proteínas variantes.

PLS y otras técnicas proporcionan otra información, más allá de la magnitud coeficiente de regresión, que se puede utilizar para clasificar residuos específicos o posiciones de los residuos. Técnicas tales como PLS y análisis de componentes principales (PCA) 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, tales como el espacio de secuencias de proteínas, la actividad empleada en esta descripción. Estos vectores latentes son funciones de las diversas dimensiones de la secuencia; es decir, los residuos individuales o posiciones de los residuos que comprenden las secuencias de proteínas de la biblioteca variante utilizada para construir el conjunto de entrenamiento. Por tanto, un vector latente comprenderá una suma de contribuciones de cada una de las posiciones de los residuos en el conjunto de entrenamiento. Algunas posiciones contribuirán con más fuerza a la dirección del vector. Éstos se manifiestan por "cargas" relativamente grandes, es decir, los coeficientes utilizados para describir el vector. Como un simple ejemplo, un conjunto de entrenamiento puede consistir en tripéptidos. El primer vector latente tendrá típicamente contribuciones de los tres residuos.

Vector 1 = al (posición de residuo 1) + a2 (posición de residuo 2) + a3 (posición de residuo 3)

Los coeficientes al, a2, y a3, son las cargas. Debido a que estos reflejan la importancia de las posiciones de residuos correspondientes a la variación en el conjunto de datos, que pueden ser utilizados para clasificar la importancia de las posiciones de residuos individuales para los propósitos de decisiones "de alternancia", como se describió anteriormente. Cargas, como los coeficientes de regresión, se pueden utilizar para clasificar los residuos en cada posición alternada. Varios parámetros describen la importancia de estas cargas. Algunos de tales

Importancia de la Variable en la Proyección (VIP) hacen uso de una matriz de carga, que se compone de las cargas de múltiples vectores latentes tomadas de un conjunto de entrenamiento. En Importancia Variable para Proyección PLS, la importancia de la i-ésima variable (por ejemplo, posición de residuo) se calcula mediante el cálculo de VIP (variable de importancia en la proyección). Para una dimensión PLS dada, a, $(VIN)_{AK}^2$ es igual al peso PLS cuadrado $(w_{ak})^2$ de una variable, multiplicada por la variabilidad de porcentaje explicada de y (variable dependiente, por ejemplo, cierta función) por esa dimensión PLS. $(VIN)_{ak}^2$ se suma sobre todas las dimensiones de PLS (componentes). VIP se calcula entonces dividiendo la suma por la variabilidad total de ciento y explicada por el modelo PLS y multiplicando por el número de variables en el modelo. Las variables con gran VIP, mayor que 1, son los más relevantes para correlacionar con una determinada función (y) - y por lo tanto mejor clasificados a los efectos de la toma de decisiones de alternancia.

Otras alternativas a la metodología anteriormente implican diferentes procedimientos para el uso de la importancia de residuo (ranking) en la determinación de qué residuos se puede alternar. En una alternativa, las posiciones de residuos de mayor puntuación se eligen para alternar. La información que se necesita en este enfoque incluye la secuencia de una mejor proteína del conjunto de entrenamiento, una mejor secuencia de PLS predicha, y una clasificación de los residuos del modelo PLS. La "mejor" proteína es un laboratorio húmedo validado "mejor" clon en el conjunto de datos (clon con la función más alta medida que aún así los modelos, es decir, caídas relativamente cerca del valor predicho de la validación cruzada PLS). El método compara cada residuo a partir de esta proteína con el residuo correspondiente de una secuencia "mejor predicha" que tiene el valor más alto de la actividad deseada. Esto se logra usando, por ejemplo, la matriz de cargas, comenzando con el residuo que tiene la carga más alta. Alternativamente, otra medida de la secuencia PLS mejor predicha tal como el valor más alto del coeficiente para cada posición de la regresión - se utiliza. Si el residuo de la carga más alta o coeficiente de regresión no está presente en la "mejor" clon, el método presenta esa posición como una posición de palanca para la biblioteca subsiguiente. El proceso se repite para varios residuos, moviéndose a través de los valores de carga sucesivamente más bajos, hasta que la biblioteca es de tamaño suficiente.

Más generalmente, una secuencia predicha por el modelo de actividad de la secuencia de tener el valor más alto (o uno de los valores más altos) de la actividad deseada se pueden utilizar de diversas maneras en la construcción de una biblioteca próxima generación, puede ser sujeto a varias mutagénesis, recombinación y / o técnicas de selección subsecuencia. Cada uno de éstos se puede realizar in vitro, in vivo, o in silico.

III. IDENTIFICACIÓN DE BIOMOLÉCULAS DE OBJETIVO CON PROPIEDADES DESEADAS Y/O PARA EVOLUCIÓN ARTIFICIAL

A. DISEÑO DE BIBLIOTECA MEDIANTE LA OPTIMIZACIÓN DE PARETO DELANTERO DE PROPIEDADES MÚLTIPLES

La presente descripción proporciona métodos que utilizan optimización de frente de Pareto para seleccionar clones para llevar a cabo futuras rondas de evolución artificial (por ejemplo, combinación aleatoria de ADN, etc.) en relación con la optimización de propiedades múltiples de polipéptido (es decir, múltiples objetivos). La optimización de frente de Pareto es un algoritmo de multi-objetivo evolutivo que mejora simultáneamente dos o más deseados objetivos.

Para ilustrar, la Figura 2 proporciona un gráfico que ilustra un frente de Pareto en una parcela de un conjunto hipotético de datos, donde la función 2 (F2) se representa gráficamente como una función de la Función 1 (FI). Cualquier problema de optimización se moldea opcionalmente como un problema de minimización, por ejemplo, mediante la inversión de la señal de la aptitud o invirtiendo la aptitud. Como se muestra en la Figura 2, por ejemplo, los ejes representan diferentes objetivos a ser simultáneamente minimizados. Las soluciones (representadas por los puntos de datos numerados) que se encuentran en el frente de Pareto representan soluciones de contrapartida que no están "dominadas" por cualquier otra solución. Estos puntos no dominados se definen por el hecho de que no existe otra solución en el hipotético conjunto de datos que es mejor (más pequeña en esta facilidad) que todas las soluciones en ambos objetivos. Por ejemplo, la solución 1 es parte del frente de Pareto, porque, aunque la solución 2 tiene un valor más pequeño para el objetivo F2, la solución 1 tiene un valor menor para el objetivo FI. En contraste, la solución de 7 no es parte de la frente de Pareto, porque al menos una solución es mejor en ambos objetivos.

La Figura 4 es un gráfico que representa ciertas etapas llevadas a cabo en una realización del método descrito en este documento de identificación de los miembros de una población de variantes de secuencia de biopolímero más adecuadas para la evolución artificial. La frase "más adecuada para la evolución artificial" se refiere a aquellos miembros de la población variante que se encuentran al menos proximal a un frente de Pareto, por ejemplo, cuando se puntúan las variantes (por ejemplo, control o seleccionado) y se representa para los objetivos deseados. Estas variantes son generalmente las más adecuadas para la evolución artificial, debido a que no están dominadas por otras variantes (o al menos la mayoría de las otras variantes) en al menos uno de los objetivos deseados.

Como se muestra en la figura 4 de **AI**, el método incluye la selección o cribado de los miembros de la población de las variantes de secuencia de biopolímero (por ejemplo, las variantes de cadenas de caracteres, etc.) para dos o más objetivos deseados para producir un conjunto de datos de la aptitud de multiobjetivos. Los objetivos deseados típicamente incluyen, por ejemplo, las propiedades estructurales y / o funcionales, tales como cualquiera de los

descritos en este documento. La población de variantes de la secuencia de biopolímeros se puede producir de acuerdo con los procedimientos de generación de diversidad descritos en este documento, luego probados respecto a actividades u otra función (es decir, objetivos). A partir de entonces, el método incluye la identificación de un frente de Pareto (por ejemplo, sustancialmente convexa, sustancialmente no convexa, etc.) en el conjunto de datos de aptitud de multi-objetivo (**A2**), y miembros de selección proximales al frente de Pareto (**A3**), con lo que se identifica los miembros de la población de las variantes de secuencia de biopolímero más adecuadas para la evolución artificial. En el contexto de la presente descripción, el "frente de Pareto" se refiere a variantes de secuencias de biopolímero que no son dominadas por otras variantes de secuencia de biopolímero en al menos uno de los dos o más objetivos deseados. En algunas realizaciones, el método incluye además la evolución de los miembros seleccionados de **A3** usando procedimientos de evolución artificial para producir variantes de secuencia de biopolímero evolucionadas. Varios procedimientos de evolución artificial que se utilizan opcionalmente para evolucionar estas variantes se describen en este documento. Al menos un paso, y en ciertos casos todos los pasos, de estos procedimientos de evolución artificial se pueden realizar in silico. Estas realizaciones también incluyen opcionalmente la repetición de los pasos **AI-A3**, utilizando variantes de secuencia de biopolímero evolucionadas como al menos algunos de los miembros de la población de variantes de secuencia de biopolímero en un paso repetido **AI**. Típicamente, al menos un paso, y algunos casos todos los pasos, de los métodos descritos en este documento se realizan en un sistema digital o basado en la web. Sistemas digitales y basados en la Web se describen en mayor detalle a continuación.

Además, para proporcionar un conjunto óptimo de soluciones entre las que elegir, los algoritmos deben intentar generalmente distribuir uniformemente o difundir las soluciones en el espacio objetivo a lo largo del frente de Pareto al máximo, ya que las soluciones agrupadas generalmente carecen de suficiente diversidad. De acuerdo con ello, los algoritmos se diseñan típicamente para pedir soluciones individuales en una población en base a tanto la aptitud a lo largo de cada objetivo y de acuerdo a su aislamiento relativo en el espacio objetivo. Este enfoque generalmente se traduce en una buena difusión de soluciones a lo largo del frente de Pareto, incluso en regiones no convexas del espacio objetivo. Frentes de Pareto no convexos se discuten más adelante. Un enfoque para la selección de soluciones basado en su diversidad relativa es la técnica de la selección basada en la región, que se describe con más detalle en, por ejemplo, Corne et al, "PESA-11: selección basada de región en la optimización evolutiva de multiobjetivo", en Actas de la Conferencia de Computación Genética y Evolutiva (GECCO-2001) Morgan Kaufmann Publishers, (2001), pp. 283-290. Selección basada en la región en general implica dividir el espacio objetivo en hyperboxes y preferentemente la selección de soluciones de hyperboxes menos pobladas. Otras técnicas para seleccionar soluciones (por ejemplo, selección de torneo binario, etc.), que son generalmente conocidos en la técnica se utilizan opcionalmente en la práctica de los métodos descritos en el presente documento.

Una ventaja significativa de la optimización de frente de Pareto es que el enfoque no para reducir el problema en cuestión a una de optimización de objetivo único (por ejemplo, por un método de suma ponderada o similares), en lugar del enfoque que proporciona un conjunto de soluciones óptimas entre las que seleccionar. A pesar de las medidas ponderadas se utilizan opcionalmente para seleccionar soluciones finales, no serán identificadas todas las soluciones a través de este enfoque, por ejemplo, si el frente de Pareto es no convexo. Por consiguiente, una simple suma ponderada de objetivos puede restringir la capacidad de un algoritmo para encontrar soluciones viables en estos casos. El problema planteado por la no-convexidad en el espacio objetivo se ilustra aún más en la Figura 3, que proporciona una gráfica que muestra una gráfica de un conjunto hipotético de datos. Como se muestra y en consonancia con la definición, el conjunto de soluciones (representado por los puntos de datos numerados) a lo largo del frente de Pareto son no dominadas. Sin embargo, la optimización clásica basada en el peso, que generalmente se conoce en la técnica, no produciría soluciones 3 y 4 para cualquier peso sobre los objetivos Fl y F2, debido a la existencia de las mejores soluciones en base a la suma ponderada. Por otra parte, si se buscaba una compensación aproximadamente igual para ambos objetivos, toda una clase de soluciones se excluiría el uso de los métodos clásicos.

Los métodos de la presente descripción incluyen diversas formas de realización para la selección de variantes de secuencia que son proximales al frente de Pareto. Por ejemplo, los métodos incluyen opcionalmente la aplicación de una o más técnicas de nichos para identificar los miembros de la población de la secuencia de variantes de biopolímeros más adecuada para la evolución artificial. Detalles adicionales relativos a diversas técnicas de nichos se proporcionan en, por ejemplo, Darwen et al. (1997) "La especiación de modularización categórica automática," IEEE Transactions on Evolutionary Computation 1 (2): 101-108, Darwen et al. (1996), "Cada método de nichos tiene su nicho: compartir la aptitud y el intercambio implícito en comparación," Proc. de resolución de problemas paralelos de la Naturaleza (PPSN) IV, Vol. 1141, Lecture Notes in Computer Science, Springer-Verlag (1996), pp.398-407, y Horn et al. (1994) "Un algoritmo genético de Pareto nichado para la optimización de multiobjetivos", en Procedimientos de la Primera Conferencia IEEE sobre Computación Evolutiva, del Congreso Mundial Computational en Computación de la IEEE, (1):82-87. En otras realizaciones, las variantes de secuencia se seleccionan por, por ejemplo, el cálculo de una suma ponderada de los dos o más deseados objetivos para al menos algunos de los miembros proximales a la frente de Pareto, y seleccionar al menos un miembro que incluye una suma más alta ponderada de otros miembros proximales a la frente de Pareto. En aún otras realizaciones, las variantes de secuencia de biopolímero se seleccionan, por ejemplo, mediante la clasificación de uno o más miembros de acuerdo con relativa proximidad a la frente de Pareto y el aislamiento relativo en el espacio de secuencias, y seleccionar al menos un miembro que clasifica más altamente que otros miembros proximales a la frente de Pareto. Las técnicas

de selección basado en la región (descritas anteriormente) también se utilizan opcionalmente para seleccionar miembros proximales a la frente de Pareto. Para ilustrar, una técnica de selección basada en la región incluye el espacio de secuencias de partición que incluye la población de las variantes de secuencia de biopolímero en una o más hyperboxes y selección de los miembros proximales a la frente de Pareto a partir de al menos uno de los hyperboxes que es menos poblada que otras regiones de la secuencia de espacio. Para ilustrar aún más, la Figura 5 es un gráfico que representa ciertas etapas llevadas a cabo en una realización de un método de identificación de los miembros de un conjunto de variantes de cadena de caracteres de biopolímero que incluyen múltiples objetivos mejorados en relación con otros miembros del conjunto de variantes de cadena de caracteres de biopolímero. Como se muestra, el método incluye la aplicación de uno o más multi-objetivos evolutivos de algoritmos para al menos una cadena de caracteres de biopolímero parentales (por ejemplo, una pluralidad de cadenas de caracteres de biopolímero parentales o similares) para producir el conjunto de variantes de cadenas de caracteres de biopolímero (**BI**) y la selección o el cribado de los miembros del conjunto de variantes de cadenas de caracteres biopolímero para dos o más objetivos deseados (**B2**). Como se muestra adicionalmente, el método incluye también el trazado de la serie de cadena de variantes de caracteres de biopolímero como una función de los dos o más objetivos deseados para producir una variante de cadena de parcela de carácter de biopolímero (por ejemplo, como se representa en la Figura 2 o 3) (**B3**), y la identificación de una frente de Pareto (por ejemplo, sustancialmente convexa, sustancialmente no convexa, etc.) en la trama de variantes de cadena de caracteres de biopolímero (**B4**), identificando con ello los miembros del conjunto de variantes de cadenas de caracteres biolpolímero que incluyen la mejora de múltiples objetivos relativos a los otros miembros de la serie de cadena de variantes de caracteres de biopolímero. El método se lleva a cabo opcionalmente de forma iterativa, por ejemplo, mediante la repetición de las etapas **B1-B4**, usando al menos un miembro del conjunto de variantes de cadena de caracteres de biopolímero como una cadena de caracteres de biopolímero parentales en un paso repetido **B1**. En algunas realizaciones, los métodos incluyen además la síntesis de polinucleótidos o polipéptidos variantes de secuencias que corresponden a los miembros del conjunto de variantes de cadenas de caracteres de biopolímeros identificados en el paso **B4**.

En realizaciones preferidas, los miembros proximales a la frente de Pareto en un análisis dado se separan maximalmente (por ejemplo, de manera sustancialmente uniforme o uniformemente distribuido) el uno del otro, por ejemplo, para mejorar la diversidad entre las soluciones identificadas, como se describe anteriormente. En otras realizaciones, la secuencia de variantes proximales a la frente de Pareto son sustancialmente distribuidas de manera desigual (por ejemplo, al azar o de manera no uniforme distribuido). Además, las parcelas variantes de cadena de caracteres de biopolímero se presentan opcionalmente como, por ejemplo, la maximización o minimización de parcelas.

Muchos objetivos deseados diferentes se criban o se seleccionan opcionalmente según estos métodos. Para ilustrar esto, cada uno de los dos o más objetivos deseados incluyen típicamente independientemente un fisicoquímico o propiedad funcional. En algunas realizaciones, los dos objetivos o más deseados incluyen, por ejemplo, las limitaciones, los valores que detallan distancia de las limitaciones que alcanzaron, un número total de restricciones satisfechos y / o un número relativo de las limitaciones satisfechas. Opcionalmente, los dos o más objetivos deseados incluyen medidas de la aptitud, la competencia u objetivos que no compitan, o similares. Además, los dos o más deseados objetivos son también opcionalmente ortogonales entre sí.

En otros aspectos, la descripción proporciona sistemas de identificación de miembros de un conjunto de variantes de cadenas de caracteres de biopolímero que incluyen múltiples objetivos mejorados en relación con otros miembros del conjunto de variantes de cadenas de caracteres de biopolímero. Los sistemas incluyen un ordenador que tiene una base de datos capaz de almacenar el conjunto de variantes de cadenas de caracteres de biopolímero. Los sistemas también incluyen el software de sistema que incluye instrucciones de lógica para la aplicación de los algoritmos evolutivos de multiobjetivo de cadenas de caracteres de biopolímero parentales para producir el conjunto de variantes de cadenas de caracteres de biopolímero, y la selección o cribado de los miembros del conjunto de variantes de cadenas de caracteres biopolímero por dos o más objetivos deseados. El software del sistema también incluye instrucciones de lógica para el trazado del conjunto de variantes de cadena de caracteres de biopolímero como una función de los dos o más objetivos deseados para producir el conjunto de variantes de secuencia de carácter de biopolímero y la identificación de una frente de Pareto en la variante de cadena de caracteres de parcela de biopolímero. Los sistemas se describen en mayor detalle a continuación.

La divulgación también proporciona un producto de programa informático que incluye un medio legible por ordenador que tiene instrucciones de lógica de aplicación de algoritmos evolutivos multi-objetivo de cadenas de caracteres de biopolímero parentales para producir un conjunto de variantes de cadenas de caracteres de biopolímero y la selección o cribado de los miembros del conjunto de variantes de cadenas de caracteres de biopolímero por dos o más objetivos deseados. Además, el producto de programa de ordenador incluye instrucciones de lógica para trazar el conjunto de variantes de cadena de caracteres de biopolímero como una función de los dos o más objetivos deseados para producir una variante de cadena de parcela de carácter de biopolímero y la identificación de una frente de Pareto en la trama de variantes de cadena de caracteres de biopolímero para identificar los miembros del conjunto de variantes de cadenas de caracteres de biopolímero que incluyen múltiples objetivos mejorados en relación con otros miembros del conjunto de variantes de caracteres de cadena de biopolímero.

Para ayudar en la selección de clones a partir de un experimento dado para desarrollar, por ejemplo, mediante los

procedimientos de evolución artificial descritos en este documento, sistemas y productos de programa de ordenador de la descripción general incluyen instrucciones lógicas que clasifican clones en términos de, por ejemplo, su proximidad a la frente de Pareto, por su relativo aislamiento, y / o similares. Esto proporciona una amplia diversidad a lo largo de la frente de Pareto con los beneficios concomitantes de tal diversidad, como se describe anteriormente. Además, los mejores clones a lo largo de la frente de Pareto más avanzada se seleccionan opcionalmente a velocidades de muestreo (por ejemplo, las concentraciones de ADN, etc.) en función de sus valores de aptitud modificados. Esto permite que los clones de las zonas menos pobladas del espacio objetivo se muestreen más a menudo, que a su vez promueve la diversidad en las siguientes rondas de evolución artificial. Una suma ponderada de las actividades después de la evolución se utiliza opcionalmente para seleccionar el "mejor" clon. Sin embargo, los investigadores han encontrado que el uso de una suma ponderada de las actividades durante la evolución resulta en un único objetivo de optimización con baja diversidad a lo largo de la frente de Pareto.

Además, las técnicas de nichos (mencionadas anteriormente) se aplican opcionalmente para seleccionar clones para el desarrollo. Por ejemplo, en optimización multi-modal de un solo objetivo, la investigación ha demostrado que los nichos pueden ser beneficiosos en ciertas circunstancias. La idea es simplemente para evolucionar artificialmente esos individuos en la población que son similares genotípicamente y que ocupan las zonas altas de la aptitud. El razonamiento es que los motivos reunidos de diferentes modos en el espacio físico no pueden conducir a una mejor función. De hecho, a menudo conducen a ruido y a interrupción. En el marco de optimizaciones de multiobjetivo, un problema del juguete simplificado puede ser simulado (por ejemplo, utilizando el modelo de NK de Kaufmann, etc.) para determinar si el sistema de nichos ayuda o dificulta la evolución a lo largo de la frente de Pareto. Véase, por ejemplo, Kauffman, Los orígenes de Orden Oxford University Press (1993) y Kaufmann y Johnsen, "Co-Evolución al borde del caos: paisajes de aptitud acoplados, Estados Asentados y avalanchas coevolutionarias". En Langton et al, Vida artificial II: Procedimientos del taller la segunda vida artificial Addison-Wesley (1992), pp. 325-369. En particular, puede depender de la robustez relativa de espacio físico de cada objetivo. Por ejemplo, los motivos que confieren, por ejemplo, la termoestabilidad pueden ser aditivos, mientras que los motivos que confieren, por ejemplo, la actividad bajo diferentes condiciones de pH puede ser competitiva y los intentos de hacer grandes saltos en el espacio de aptitud de multi-objetivo puede dar lugar a altas tasas de muertos.

B. EVOLUCIÓN IN SILICO

La presente descripción incluye métodos de optimización de construcción de la biblioteca a través de la evolución in silico de las bibliotecas que utilizan algoritmos de búsqueda evolutivos, incluyendo algoritmos genéticos y métodos de Monte Carlo, que se describen en este documento. Estos métodos maximizan el éxito in vivo y / o en la evolución in vitro de esencialmente cualquier material genético, incluidos genes, operones, caminos, promotores, elementos reguladores, genomas, o similares.

Más específicamente, la Figura 6 proporciona un diagrama que representa ciertos pasos llevados a cabo en una realización del método para la evolución de las bibliotecas para la evolución dirigida en la que la biblioteca (L) es la unidad de la evolución en el algoritmo. Cada biblioteca se describe mediante parámetros tales como la diversidad de secuencia, el método de recombinación, las condiciones experimentales, y / o similares. Los parámetros adicionales se describen en el presente documento. Los parámetros se cambian normalmente o no evolucionan durante el proceso de evolución. Como se muestra en **C2**, los métodos incluyen la provisión de una población de bibliotecas (por ejemplo, una población inicial de bibliotecas (**C1**)), tales como las poblaciones de variantes de cadena de caracteres de biopolímero. El algoritmo incluye un conjunto de operadores (0) que opera en la unidad L para producir una nueva población de bibliotecas (**C3**). Por ejemplo, las operaciones incluyen la adición y eliminación de la diversidad, el cambio de las tasas de recombinación y frecuencias y / o similares. Detalles adicionales con respecto a los operadores que se utilizan opcionalmente en estos métodos se proporcionan en este documento. En particular, el operador actúa sobre una población de bibliotecas para crear la próxima generación de la población. Como se muestra en **C4**, se selecciona entonces esta nueva generación para la aptitud (F) para producir una población ajustadora de bibliotecas (**C5**) y este proceso se repite (**C6**). Este algoritmo evolutivo se detuvo cuando normalmente las características deseadas (por ejemplo, niveles de aptitud) se cumplen para las bibliotecas. Opcionalmente, el proceso de selección implica el diseño de oligonucleótidos utilizando algoritmos para facilitar la identificación de las secuencias de datos correspondientes a polímeros biológicos y enumerando / simulando el resultado de un experimento, seguido de estimación *in silico* de las actividades de los clones. Cada biblioteca es entonces típicamente caracterizado por una función de aptitud que consiste en determinar, por ejemplo, la actividad de los clones, la desviación estándar de las actividades de los clones, la diversidad genética entre clones, la simplicidad experimental de la biblioteca, etc. Las actividades de los clones también pueden caracterizarse por redes neuronales, PCA u otras herramientas de predicción o mediante la compatibilidad estructural, simulación de dinámica y otros métodos biofísicos y / o por otras técnicas se describe en el presente documento.

Para ilustrar aún más estos aspectos de la divulgación, la Figura 7 proporciona un gráfico que muestra ciertas etapas llevadas a cabo en una realización de un método de producción de una población más en forma de bibliotecas de cadenas de caracteres que utiliza varios operadores. Al menos un paso, y en ciertos casos todos los pasos, del método es / son realizadas típicamente *in silico*, por ejemplo, en un sistema digital descrito en este documento. Como se muestra, el paso **DI** incluye la aplicación de uno o más operadores a una población inicial de las bibliotecas de cadenas de caracteres para producir una población evolucionada de las bibliotecas de cadenas de

caracteres. Típicamente, uno o más caracteres de cadenas en la población inicial de las bibliotecas de cadenas de caracteres corresponden a uno o más polinucleótidos o uno o más polipéptidos. Después de asignar un nivel de aptitud (por ejemplo, la selección o chequeo por, por ejemplo, propiedades estructurales deseadas, las propiedades deseadas funcionales, y / o similares) a los miembros de la población evolucionada de las bibliotecas de cadenas de caracteres (**D2**), el método incluye la selección de los miembros de la población evolucionada de bibliotecas de cadenas de caracteres con niveles más altos de la aptitud que otros miembros de la población para producir una población más apta de bibliotecas de cadenas de caracteres (**D3**). Los procedimientos incluyen además la repetición de los pasos **D1-D3**, utilizando la población más apta del carácter de cadena de bibliotecas como la población inicial de las bibliotecas de cadenas de caracteres en un paso repetido **Dl**, por ejemplo, hasta un nivel deseado de la aptitud se alcanza en al menos una cadena de caracteres de biblioteca.

En ciertas realizaciones, el paso **Dl** incluye (i) la provisión de conjuntos de subcadenas degeneradas basadas en la población inicial de miembros de bibliotecas de cadenas de caracteres, (ii) la recombinación de los conjuntos de subcadenas degeneradas para producir cadenas variadas de caracteres deseadas sistemáticamente, y (iii) la estimación de una o más actividades de las cadenas de caracteres variadas sistemáticamente deseadas para producir la población evolucionada de las bibliotecas de cadenas de caracteres. En algunas realizaciones, uno o más miembros de la población inicial de las bibliotecas de cadenas de caracteres se definen por un algoritmo que toma uno o más parámetros, que los parámetros evolucionan durante el paso **Dl**. Los parámetros de ejemplo incluyen, por ejemplo, la diversidad de cadena de caracteres, el método de evolución modelado utilizado, las condiciones experimentales modeladas utilizadas, modelización PCA, modelización PLS, matrices de mutación, la importancia relativa de, por ejemplo, cadenas de caracteres individuales o bibliotecas, sistemas de puntuación para algunos o todos los parámetros utilizados, y / o similares. La población inicial de bibliotecas de cadenas de caracteres incluye generalmente entre dos y unos 10^5 bibliotecas. Además, cada biblioteca de cadena de caracteres de la población inicial de las bibliotecas de cadenas de caracteres incluye típicamente entre aproximadamente dos y aproximadamente 10^5 miembros.

Muchos operadores diferentes se utilizan opcionalmente en la práctica de estos métodos. Estos incluyen, por ejemplo, una mutación de uno o más miembros de cadena de caracteres de bibliotecas, una multiplicación de uno o más miembros de la cadena de caracteres de bibliotecas, una fragmentación de uno o más miembros de las bibliotecas de cadenas de caracteres, un cruce entre los miembros de las bibliotecas de cadenas de caracteres, una ligación de uno o más miembros de las bibliotecas de cadenas de caracteres o subcadenas de uno o más miembros de las bibliotecas de cadenas de caracteres, un cálculo de elitismo, un cálculo de la secuencia de homología o similitud de secuencia de cadenas de caracteres alineados, un uso recursivo de uno o más operadores genéticos para la evolución de uno o más miembros de las bibliotecas de cadenas de caracteres, una aplicación de un operador de aleatoriedad a uno o más miembros de las bibliotecas de cadenas de caracteres, una mutación por deleción de uno o más miembros de las bibliotecas de cadenas de caracteres, una mutación de inserción en uno o más miembros de las bibliotecas de cadena de caracteres, resta de uno o más miembros de las bibliotecas de cadenas de caracteres, selección de uno o más miembros de las bibliotecas de cadenas de caracteres con las actividades deseadas, la muerte de uno o más miembros de las bibliotecas de cadenas de caracteres, o similares. Véase, por ejemplo, WO 00/42560; WO 01/75767. Los operadores generalmente se incluyen como componentes de algoritmos evolutivos de búsqueda. La búsqueda evolutiva preferida de algoritmos incluyen algoritmos genéticos, algoritmos de Monte Carlo y / o similares, los cuales también se describen adicionalmente en este documento.

Los niveles de aptitud se suelen asignar a cada miembro de la población evolucionada de las bibliotecas de cadenas de caracteres, utilizando funciones de aptitud. Funciones ejemplares de la aptitud opcionalmente incluyen, por ejemplo, la determinación de actividades medias de los miembros de cada biblioteca de cadena de caracteres, la determinación de las desviaciones estándares de las actividades de los miembros de cada biblioteca de cadena de caracteres, la determinación de niveles de diversidad de cadena de caracteres entre los miembros de cada biblioteca de cadena de caracteres, el modelado de una simplicidad experimental de cada biblioteca de cadena de caracteres, la determinación de un nivel de confianza en valores medidos o predichos, y / o similares. En realizaciones preferidas, las actividades de los miembros se determinan utilizando técnicas de análisis multivariante y / o técnicas de análisis biofísicas. Por ejemplo, las técnicas de análisis de multivariante opcionalmente incluyen, por ejemplo, los análisis de las técnicas de entrenamiento de redes neuronales, componentes principales, análisis de mínimos cuadrados parciales, y / o similares. Técnicas de análisis biofísicas típicas incluyen uno o más de, por ejemplo, los análisis de compatibilidad estructural, simulaciones de dinámica, los análisis de hidrofobicidad, los análisis de solubilidad, los análisis de inmunogenicidad, ensayos de unión, caracterizaciones enzimáticas, o similares. El análisis multivariado y análisis biofísicos se describen con más detalle en el presente documento.

Los miembros de la población en forma de bibliotecas de cadenas de caracteres en general corresponden a los polinucleótidos o polipéptidos. Aunque las etapas de estos métodos se llevan a cabo típicamente *in silico* (por ejemplo, utilizando un sistema digital, un sistema basado en la web, etc.), los métodos incluyen además opcionalmente la síntesis de, por ejemplo, uno o más de los polinucleótidos o polipéptidos que corresponden a una o más miembros de la población en forma de bibliotecas de cadenas de caracteres para producir polinucleótidos o polipéptidos sintetizados. Además, los métodos también opcionalmente incluyen, por ejemplo, la selección o cribado de los polinucleótidos o polipéptidos sintetizados por al menos una propiedad deseada para producir filtrado o polinucleótidos o polipéptidos seleccionados. Típicamente, los polinucleótidos o polipéptidos sintetizados se

seleccionan in vitro o in vivo. Diversas técnicas de detección utilizadas en la práctica de estos métodos se describen en el presente documento. Los métodos incluyen opcionalmente además el sometimiento de los polinucleótidos o polipéptidos seleccionados o cribados para uno o más procedimientos artificiales de evolución. Al menos una etapa de los uno o más procedimientos de evolución artificial se realiza opcionalmente *in silico*, por ejemplo, el uso de representaciones de cadena de caracteres de los polinucleótidos o polipéptidos.

En otro aspecto, la descripción se refiere a un sistema para producir una población más en forma de bibliotecas de cadenas de caracteres. El sistema incluye (a) al menos un ordenador que incluye una base de datos capaz de almacenar al menos una población de las bibliotecas de cadenas de caracteres, y (b) el software del sistema incluye una o más instrucciones de lógica. Las instrucciones de lógica son típicamente de, por ejemplo, (i) la aplicación de uno o más de los operadores a una población inicial de bibliotecas de cadenas de caracteres para producir una población evolucionada de las bibliotecas de cadenas de caracteres, (ii) la asignación de un nivel de aptitud a por lo menos un miembro de la población evolucionada de las bibliotecas de cadenas de caracteres, (iii) seleccionar uno o más miembros de la población evolucionada de las bibliotecas de cadenas de caracteres con altos niveles de aptitud que otros miembros de la población evolucionada de las bibliotecas de cadenas de caracteres para producir la población en forma de bibliotecas de cadenas de caracteres, y (iv) repetir las etapas (i) - (iii) usando el población instaladora de las bibliotecas de cadenas de caracteres como la población inicial de las bibliotecas de cadenas de caracteres en un paso repetido (i). El sistema normalmente incluye además un polinucleótido o un dispositivo de síntesis de polipéptidos capaces de síntesis de polinucleótidos o polipéptidos que corresponden a los miembros de la población en forma de bibliotecas de cadenas de caracteres. Los sistemas se describen en mayor detalle a continuación.

La divulgación también proporciona un producto de programa informático que incluye un medio legible por ordenador que tiene una o más instrucciones lógicas para (a) la aplicación de uno o más operadores a una población inicial de las bibliotecas de cadenas de caracteres para producir una población evolucionada de las bibliotecas de cadenas de caracteres, y (b) la asignación de un nivel de aptitud a por lo menos un miembro de la población evolucionada de las bibliotecas de cadenas de caracteres. El producto de programa de ordenador también incluye instrucciones lógicas para (c) seleccionar uno o más miembros de la población evolucionada de las bibliotecas de cadenas de caracteres con niveles más altos de la aptitud que otros miembros de la población evolucionada de las bibliotecas de cadenas de caracteres para producir la población en forma de cadena de caracteres de bibliotecas y (d) repetir las etapas (a)-(c) utilizando la población en forma de bibliotecas de cadenas de caracteres como la población inicial de las bibliotecas de cadenas de caracteres en un paso repetido (a).

C. FABRICACIÓN DE BIBLIOTECAS DE MODELOS HEURÍSTICAMENTE DERIVADOS

La siguiente discusión complementa el aspecto descrito anteriormente de la divulgación presentada en la Figura 1. También se presentan algunas realizaciones alternativas y elabora en algunos conceptos previamente introducidos. No limita la discusión anterior.

Tal como se describe en el presente documento, teniendo acceso a los conjuntos de datos de secuencias sistemáticamente variadas con actividades medidas permite la generación de varios modelos. Esta descripción ilustra cómo implementar estos modelos en la construcción de bibliotecas preferidas. Aunque otras técnicas de modelado, muchos de los cuales se describen en el presente documento, están opcionalmente también utilizados para construir/puntuar bibliotecas, a los modelos PLS se hace hincapié en esta sección para mayor claridad. En particular, una de las alternativas para decidir sobre la secuencia de espacio de búsqueda consiste en aislar las cargas (por ejemplo, las relaciones de función) de cada residuo de aminoácido en una alineación determinada. Por ejemplo, las cargas típicamente se encuentran almacenadas como una matriz en el modelo generado, por ejemplo, por cualquier herramienta de modelado PLS estándar y se pueden recuperar, por ejemplo, a partir de una matriz File_Name.loads.

En general, la importancia para cada residuo y mejor, por ejemplo, 5% de pares de residuos (definidos como productos cruzados en la matriz) se determina opcionalmente utilizando PLS o similares, y la importancia relativa se da como carga (si uno de los componentes se utiliza), coeficiente de regresión, VIP (importancia variable para la proyección), etc. Opcionalmente, las cargas se clasifican posteriormente, por ejemplo, de acuerdo con el valor numérico. El aminoácido preferido en cada posición en la proteína particular que tiene dos o más aminoácidos opcionales será determinado por el correspondiente aminoácido que tiene la carga más alta, el coeficiente de regresión, VIP, etc. Un clon de "héroe" que tiene en teoría la mejor secuencia (es decir, codifica la opción de aminoácidos que tiene la carga más alta en cada posición) se determina de este modo. Además, para los modelos de generación de más de una variable latente, los coeficientes de regresión o parámetros similares pueden también ser utilizados.

Como se ha explicado, estos enfoques pueden incluir inicialmente la identificación del clon de laboratorio húmedo validado como "mejor" en un conjunto de datos en particular, que suele ser el clon con la función más alta medida que todavía modela bien (es decir, cae relativamente cerca del valor predicho de la validación cruzada PLS). Cada residuo en el mejor clon típicamente se compararon con los de la matriz de cargas, por ejemplo, comenzando con el residuo que tiene la carga más alta. Si el residuo de la carga más alta no está presente en el "mejor" clon, esa

posición se introduce como una palanca en la biblioteca subsiguiente. En algunas formas de realización, los residuos para alternar están determinados por la clasificación de cada residuo mediante el aumento de VIP y omitiendo las que están bien caracterizados en el modelo (es decir, existen en el conjunto de datos como muchos casos, y son sistemáticamente variadas). Esto puede ser más fácil de hacer mediante la retención de sólo aquellos que se producen como única (y el doble si el conjunto de datos es lo suficientemente grande) de los casos. Una biblioteca de dos por tanto codificaría el clon "héroe" y alternancia del residuo que tiene el VIP más cercano a cero y sólo está presente en un solo caso en el conjunto de datos. Una biblioteca de 4 (2^2) sería alternar los dos residuos de VIP más bajos con casos individuales, etc. Estos procesos se repiten hasta que la biblioteca alcanza un tamaño seleccionado o suficiente. Cada diversidad añadida representada por una palanca, duplica el tamaño de la biblioteca de tal manera que 10 posiciones iguala aproximadamente 1.000 clones (1.024), 13 posiciones igualan aproximadamente 10.000 clones (8192), 20 posiciones equivalen a aproximadamente 1.000.000 de clones (1.048.576), etc. El tamaño de la biblioteca apropiada depende de factores tales como el costo de la pantalla, la aspereza del paisaje, el muestreo de porcentaje preferido del espacio, y similares. Opcionalmente, restos que tienen cargas pequeñas son alternadas, por ejemplo, para buscar el espacio local que rodea a un clon "mejor" ya validado. Una opción adicional incluye comenzar con un clon de media que modela bien y alternar las altas cargas, por ejemplo, para explorar el espacio más grande en la búsqueda de las colinas de actividad previamente omitidas de la toma de muestras. Este tipo de biblioteca es generalmente más relevante en las primeras rondas, ya que genera una imagen más refinada para las rondas posteriores. Como filtro adicional, se puede omitir los residuos que se derivan originalmente de la diversidad no natural. El motivo es que la diversidad natural existente tiene una mayor probabilidad de funcionalidad de codificación que la diversidad que ocurre al azar, que puede o no ser cierto.

Para ilustrar aún más, la Figura 8 es un gráfico que muestra ciertos pasos llevados a cabo en una realización de un método de selección de posiciones de aminoácidos en una variante de polipéptido de evolucionar artificialmente, que las etapas se realizan típicamente en un sistema digital o basado en la web. Como se muestra, los métodos incluyen la provisión de una población de variantes de polipéptidos (**EI**) y puntuación (por ejemplo, *in silico*) miembros de la población de variantes del polipéptido (por ejemplo, variantes de cadenas de caracteres, etc.) para una o más propiedades deseadas (propiedades por ejemplo, estructurales y / o funcionales) para producir un polipéptido conjunto de datos de la variante (**E2**). La población de variantes de polipéptido generalmente se proporciona por uno o más procedimientos artificiales de evolución. Además, al menos una etapa (y con frecuencia más) de los procedimientos de evolución artificial se realiza típicamente *in silico*. Las poblaciones de variantes de polipéptidos incluyen típicamente, por ejemplo, entre aproximadamente dos y alrededor de 10^6 miembros. En realizaciones preferidas, los miembros de la población de variantes de polipéptidos son secuencias variadas sistemáticamente.

Los métodos incluyen además la correlación de los aminoácidos en las posiciones de aminoácidos en el polipéptido variantes con las una o más propiedades deseadas utilizando el conjunto de datos de variantes de polipéptidos para producir una matriz de cargas (por ejemplo, una matriz cualitativa (por ejemplo, incluyendo identidades de aminoácidos, etc.), una matriz cuantitativa (por ejemplo, incluyendo propiedades fisicoquímicas, tales como medidas de hidrofobicidad, etc.), una matriz categórica (por ejemplo, si los aminoácidos se pueden alojar, voluminosos, etc.), y / o similares), por ejemplo, que representan las contribuciones de aminoácidos a las propiedades deseadas (E3). Por ejemplo, si dos secuencias de polipéptidos son idénticos excepto por un único residuo de aminoácido, y las secuencias tienen actividades diferentes, entonces toda la diferencia en la función típicamente se supone que correlaciona solamente con la diferencia de un aminoácido. En consecuencia, esencialmente cualquier modo que la importancia relativa para una variable dada hacia un parámetro Y funcional se puede marcar se utiliza opcionalmente en estos métodos. Para ilustrar, la matriz se basa opcionalmente en algoritmos basados en la regresión, por ejemplo, PLS, los coeficientes de regresión, VIP (importancia variable para la proyección) (un algoritmo preferido), MLR (regresión lineal múltiple), ILS (inverso menos cuadrados), PCR (principal componente de regresión), y / o similares. Alternativas adicionales incluyen basar la matriz de cargas en algoritmos basados en patrones, tales como redes neuronales, CART (árboles de clasificación y regresión), MARS (multivariados splines de adaptación de regresión), y / o similares. También los métodos suelen incluir ordenación de las entradas en la matriz de cargas, por ejemplo, de acuerdo con valor numérico, etc.

Como se muestra en el paso **E4**, los métodos también incluyen la identificación de una o más diferencias de aminoácidos entre al menos un miembro seleccionado de entre la población de variantes de polipéptidos y las entradas correspondientes en la matriz de cargas, seleccionando de ese modo las posiciones de aminoácidos en la variante de polipéptido de evolucionar artificialmente (por ejemplo, alternar con residuos de aminoácidos variables). Por ejemplo, la solución preferida es escoger un miembro que es "mejor" o de puntuación más alta en la función preferida o conjunto de funciones (por ejemplo, el tiempo que se ajusta al modelo razonablemente bien) y escoger residuos para evolucionar en ese miembro. Típicamente, entre aproximadamente dos y aproximadamente 100 posiciones de aminoácidos en la variante de polipéptido se seleccionan para evolucionar artificialmente. Opcionalmente, se seleccionan todas las posiciones de aminoácidos en una variante dada. En ciertas realizaciones, al menos un miembro seleccionado de la población de variantes de polipéptidos en E4 incluye un miembro de puntuación más alta de **E2**. Los métodos típicamente incluyen la evolución artificial de uno o más de las posiciones de los aminoácidos seleccionados en **E4** para producir una biblioteca de polipéptidos evolucionada. Además, los métodos también incluyen la repetición opcional de **E1-E4**, utilizando la biblioteca de polipéptidos evolucionada como la población de variantes de polipéptido en una **EI** repetida. Bibliotecas de polipéptidos evolucionadas incluyen

opcionalmente bibliotecas físicaz o computacionales. Bibliotecas físicas típicamente incluyen, por ejemplo, entre aproximadamente dos y alrededor de 10^6 miembros. Por el contrario, las bibliotecas de cálculo típicamente incluyen, por ejemplo, entre aproximadamente dos y aproximadamente 10^{20} miembros.

Como se ha dicho anteriormente, en realizaciones preferidas, matrices de carga se generan a partir de conjuntos de datos de variante de polipéptido que usan varias técnicas de modelado derivadas heurísticamente, incluyendo algoritmos basados en la regresión, el patrón de base de algoritmos, y / o similares, algoritmos basados en la regresión de ejemplo incluyen, por ejemplo, regresión de cuadrados parciales mínimos, regresión lineal múltiple, regresión de mínimos cuadrados inversa, regresión de componentes principales, la variable de importancia para la proyección, etc. algoritmos basados en patrones de ejemplo incluyen, por ejemplo, redes neurales, la clasificación y árboles de regresión, splines de regresión multivariante de adaptación, y / o similares. En ciertas realizaciones preferidas, el **E3** incluye la generación de un modelo de mínimos cuadrados parciales a partir de los datos de la variante de polipéptido establecidos para producir la matriz de cargas. El modelo de mínimos cuadrados parciales normalmente genera más de una variable latente. Los métodos también incluyen típicamente aún más el uso de los coeficientes de regresión.

En realizaciones preferidas, la etapa **E4** incluye la comparación de una o más posiciones de aminoácidos en al menos un miembro con una o más posiciones de aminoácidos correspondientes de las cargas de matriz para identificar al menos un aminoácido en la matriz de cargas que está ausente en el miembro para seleccionar las posiciones de aminoácidos en la variante de polipéptido a evolucionar artificialmente. Generalmente, cada posición de aminoácido en al menos un miembro se compara con cada posición de aminoácido correspondiente de la matriz de cargas. Posiciones de aminoácidos seleccionadas están evolucionadas opcionalmente y artificialmente mediante la sustitución de uno o más aminoácidos correspondientes de la matriz de cargas. Además, el miembro seleccionado de la población de variantes de polipéptido incluye típicamente un miembro de puntuación más alta (por ejemplo, el miembro de puntuación más alta) del conjunto de datos de variante de polipéptido que los otros miembros del conjunto de datos variante de polipéptido. Por ejemplo, el miembro de puntuación más alta es típicamente proximal a una puntuación predicha en una validación cruzada parcial de mínimos cuadrados. Las posiciones de los aminoácidos de la matriz de cargas que incluyen cargas más altas típicamente se comparan antes de las posiciones de los aminoácidos de la matriz de cargas que incluyen cargas más bajas. Opcionalmente, las posiciones de aminoácidos de la matriz de cargas que incluyen cargas más bajas se comparan antes de las posiciones de aminoácidos a partir de la matriz de cargas que incluyen cargas más altas. En algunas realizaciones, el miembro seleccionado de la población de variantes de polipéptidos incluye un miembro de puntuación sustancialmente media del conjunto de datos de variante de polipéptido. En estas realizaciones, las posiciones de aminoácidos a partir de la matriz de cargas que incluyen cargas más altas típicamente se comparan con anterioridad a las posiciones de aminoácidos a partir de la matriz de cargas que incluyen cargas más bajas.

La Figura 9 es un gráfico que muestra ciertas etapas llevadas a cabo en otra realización de estos métodos de selección de las posiciones de aminoácidos en una variante de polipéptido de evolución artificial. Como se muestra, el método incluye la provisión de una población de polipéptido variante (**F1**), y miembros de puntuación de la población de variantes de polipéptidos de una o más propiedades deseadas para producir un conjunto de datos de variante de polipéptido (**F2**). En la etapa **F3**, un modelo de mínimos cuadrados parciales se genera a partir del conjunto de datos de variante de polipéptido, cuyo modelo de mínimos cuadrados parcial se correlaciona posiciones de aminoácidos en las variantes de polipéptidos con una o más propiedades deseadas para producir una matriz de cargas. Los métodos también incluyen la identificación de una o más diferencias de aminoácidos entre al menos un miembro seleccionado de la población de variantes de polipéptidos y la matriz de cargas desde el modelo de mínimos cuadrados parciales, seleccionando de este modo las posiciones de aminoácidos en la variante de polipéptido de evolución artificial (**F4**).

La descripción también proporciona un sistema para seleccionar posiciones de aminoácido en una variante de polipéptido de cadena de caracteres que evolucionan de forma artificial. El sistema incluye (a) un equipo que incluye una base de datos capaz de almacenar al menos una población de variantes de cadena de caracteres de polipéptido, y (b) el software del sistema. El software del sistema incluye una o más instrucciones lógicas para (i) proporcionar una o más poblaciones de variantes de cadenas de caracteres de polipéptido, y (ii) los miembros de puntuación de las una o más poblaciones de variantes de cadenas de caracteres de polipéptido para una o más propiedades deseadas para producir un conjunto de datos de variante de cadena de caracteres de polipéptido. El software también incluye instrucciones para (iii) correlacionar los aminoácidos en las posiciones de aminoácidos en las variantes de cadena de caracteres de polipéptido con las una o más propiedades deseadas, utilizando los datos de la variante de cadena de caracteres de polipéptido establecidos para producir unas cargas matriz que representan las contribuciones de aminoácidos a las una o más propiedades deseadas, y (iv) identificar una o más diferencias de aminoácidos entre al menos un miembro seleccionado entre una o más poblaciones de variantes de cadenas de caracteres polipéptido y las entradas correspondientes en la matriz de cargas. Detalles adicionales relativos a diversos aspectos de los sistemas de la presente descripción se proporcionan a continuación.

Además, la descripción se refiere a un producto de programa de ordenador para seleccionar las posiciones de aminoácidos en una variante de cadena de caracteres de polipéptido de evolución artificial. El producto de programa de ordenador incluye un medio legible por ordenador que tiene una o más instrucciones lógicas para (a)

proporcionar una o más poblaciones de variantes de cadenas de caracteres de polipéptidos, y (b) los miembros de puntuación de las una o más poblaciones de variantes de cadenas de caracteres de polipéptido para una o más propiedades deseadas para producir un conjunto de datos de variante de cadena de caracteres de polipéptido. El programa también incluye instrucciones para (c) correlacionar los aminoácidos en las posiciones de aminoácidos en las variantes de cadenas de caracteres de polipéptido con las una o más propiedades deseadas mediante los datos de variante de cadena de caracteres de polipéptido que se define para producir cargas de un matriz que representa las contribuciones de aminoácidos a las una o más propiedades deseadas, y (d) identificar una o más diferencias de aminoácidos entre al menos un miembro seleccionado de entre una o más poblaciones de variantes de cadena de caracteres de polipéptido y las entradas correspondientes en la matriz de cargas.

D. EL USO DE PRODUCTOS CRUZADOS en MODELOS DERIVADOS HEURÍSTICAMENTE PARA LA EXPLORACIÓN DE SECUENCIA DEL ESPACIO

Interacciones (por ejemplo, de segundo orden, tercer orden, etc.) entre los residuos de aminoácidos son importantes para relaciones de actividad de secuencia de proteína (función) (PSAR (PSFR)). Otro aspecto de la descripción consiste en calcular términos de productos cruzados, es decir, los residuos de co-variable, entre varias columnas correspondientes a las posiciones de aminoácidos de los residuos de ácido en una matriz. Una descripción detallada de los fenómenos de covariación se proporciona en los ejemplos a continuación. Los términos de productos cruzados se añaden típicamente a los términos lineales, que corresponden a los residuos de aminoácidos, y una matriz predictor X ampliado se genera. Modelos derivados heurísticamente se generan con la matriz predictora ampliada para identificar términos cruzados importantes junto con los términos lineales. Este producto cruzado y la información de término lineal son entonces típicamente utilizados en la construcción de bibliotecas posteriores. Por ejemplo, dos residuos de aminoácidos por sí solos no pueden ser importantes, por ejemplo, tal como se manifiesta por los pesos de los términos lineales en el modelado PLS, pero su término de producto cruzado puede ser importante. En consecuencia, las posiciones de aminoácidos correspondientes pueden ser buenos candidatos para la exploración en las siguientes rondas de evolución artificial para garantizar la óptima búsqueda de espacios de secuencia.

Para ilustrar aún más, la Figura 10 es un gráfico que muestra ciertas etapas llevadas a cabo en una realización de un método de identificación de aminoácidos en polipéptidos que son importantes para una relación de secuencia-actividad de polipéptido. Como se muestra en **GI**, los métodos incluyen la provisión de una matriz predictora X que incluye un conjunto de datos correspondientes a un conjunto de variantes de secuencia de polipéptidos en las que al menos un subconjunto del conjunto de variantes de la secuencia de polipéptidos incluyen una o más actividades medidas. El conjunto de variantes de secuencia de polipéptidos típicamente incluye, por ejemplo, un conjunto de secuencias de polipéptidos variadas de manera sistemática o similares, por ejemplo, producido por uno o más procedimientos de generación de diversidad o de evolución artificial, tales como cualquiera de los descritos en este documento. Como se muestra además en **G2**, los métodos también incluyen el cálculo de uno o más términos de productos cruzados entre columnas de la matriz X predictora. Cada entrada de columna corresponde a un aminoácido de una variante de secuencia de polipéptido del conjunto de variantes de secuencia de polipéptido. Además, los métodos también incluyen la adición de al menos uno de los uno o más términos de productos cruzados calculados en el paso **G2** a uno o más términos lineales de la matriz X predictora para producir una matriz X predictora expandida (**G3**).Términos de productos cruzados identifican aminoácidos de covariación en los polipéptidos, mientras que los términos lineales corresponden a los aminoácidos en las variantes de secuencia de polipéptido. A partir de entonces, los métodos incluyen la generación de un modelo con la matriz X predictora expandida para identificar importantes términos de productos cruzados y / o términos lineales, identificando de este modo los aminoácidos en los polipéptidos que son importantes para la relación secuencia-actividad de polipéptido (**G4**).

Opcionalmente, los modelos heurísticamente derivados se producen utilizando uno o más algoritmos basados en regresión seleccionados de, por ejemplo, una regresión de mínimos cuadrados parciales, una regresión lineal múltiple, una regresión inversa de mínimos cuadrados, un director de regresión de componentes, una variable de importancia para la proyección o similar. Como opción adicional, el modelo se produce usando uno o más algoritmos basados en el patrón seleccionado de, por ejemplo, una red neural, un árbol de clasificación y regresión, un spline de regresión multivariante adaptativa, o similares.

Típicamente, los términos importantes de productos cruzados y / o términos lineales identificados en **G4** se utilizan para diseñar una o más bibliotecas de polipéptidos. Como se mencionó, en ciertos aspectos, dos o más términos lineales de forma individual pueden incluir condiciones de poca importancia para la relación de secuencia-actividad de polipéptido. Sin embargo, los términos de productos cruzados calculados a partir de los mismos dos o más términos lineales pueden ser identificados como importantes para la relación de secuencia-actividad de polipéptido. Los términos de productos cruzados típicamente corresponden a las interacciones entre dos o varios aminoácidos en variantes de secuencia de polipéptido. Por ejemplo, las interacciones incluyen, p.ej., las interacciones secundarias o terciarias, interacciones directas, interacciones indirectas, las interacciones físico-químicas, las interacciones debidas a intermediarios de plegamiento, los efectos de la traducción, y / o similares. La información de secuencia-actividad derivada de análisis de la covariación (es decir, términos de productos cruzados) se puede utilizar en un método para la caracterización de la covariación en una biblioteca de polipéptido por:

(A) la identificación de variación de residuos de aminoácidos en una población de cadena de caracteres que representa una población de polipéptidos parentales homólogas;

(B) la identificación de residuos de aminoácidos en la población de cadena de caracteres que covarían uno con el otro para producir un conjunto de datos de covariación parental;

(C) la provisión de un conjunto de oligonucleótidos sintéticos solapantes que comprenden miembros que codifican uno o más residuos de aminoácidos que covarían identificadas en la población de cadena de caracteres,

en el que cada uno de los oligonucleótidos sintéticos codifica al menos un miembro de un conjunto de residuos de aminoácidos que covarían entre sí;

(D) recombinación de los oligonucleótidos sintéticos solapantes para producir un conjunto de polinucleótidos recombinados que codifican los polipéptidos de la progenie de parentales homólogos,

(E) expresa al menos un subconjunto del conjunto de recombinado de polinucleótidos para producir un conjunto de polipéptidos de progenie;

(F) la selección o cribado de al menos un subconjunto de los polipéptidos de la progenie de una propiedad deseada;

(G) la secuenciación de uno o más polipéptidos de la progenie, o uno o más polinucleótidos recombinados que codifican uno o más polipéptidos de progenie, que comprenden la propiedad deseada para producir un conjunto de datos de secuencias de progenie;

(H) la identificación de uno o más pares de residuos de aminoácidos en el conjunto de datos de secuencia progenie que covarían uno con el otro para producir un conjunto de datos de covariación de progenie; y

(I) la identificación de las diferencias entre los conjuntos de datos de covariación de progenie y parentales, caracterizando así la covariación en la población de polipéptidos homólogos.

Estos aspectos de la divulgación también se incorporan en un sistema para la identificación de los aminoácidos en los polipéptidos que son importantes para una relación de actividad de secuencia de polipéptido. El sistema incluye (a) un equipo que incluye una base de datos capaz de almacenar al menos una población de bibliotecas de cadenas de caracteres, y (b) el software del sistema. El software del sistema incluye una o más instrucciones lógicas para (i) proporcionar una matriz de predictor X que incluye un conjunto de datos correspondiente a un conjunto de secuencia de polipéptidos variantes en las que al menos un subconjunto del conjunto de variantes de la secuencia de polipéptidos incluye una o más actividades medidas y (ii) calcular uno o más términos de productos cruzados entre columnas de la matriz predictora X en la que cada entrada de la columna corresponde a un aminoácido de una variante de secuencia de polipéptido del conjunto de variantes de la secuencia de polipéptidos. El software también incluye instrucciones para (iii) añadir al menos uno de los uno o más términos de producto cruzado calculado en la etapa (ii) a uno o más términos lineales de la matriz predictora X para producir una matriz predictora X expandida, y (iv) la generación de un modelo con la matriz predictora X ampliada para identificar importantes términos de productos cruzados y / o términos lineales. Detalles adicionales con respecto a los sistemas de la descripción se describen a continuación.

La divulgación también proporciona un producto de programa de ordenador para la identificación de los aminoácidos en los polipéptidos que son importantes para una relación de secuencia de actividad de polipéptido. El producto de programa de ordenador incluye un medio legible por ordenador que tiene una o más instrucciones lógicas para (a) proporcionar una matriz predictora X que incluye un conjunto de datos correspondiente a un conjunto de secuencia de polipéptidos variantes en las que al menos un subconjunto del conjunto de variantes de secuencia de polipéptido incluyen una o más actividades medidas, y (b) el cálculo de uno o más términos de productos cruzados entre columnas de la matriz predictora X en el que cada entrada de la columna corresponde a un aminoácido de una variante de secuencia de polipéptido del conjunto de variantes de secuencia de polipéptidos. El programa también incluye instrucciones para (c) añadir al menos uno de los uno o más términos de productos cruzados calculados en (b) a uno o más términos lineales de la matriz predictora X para producir una matriz predictora X expandida, y (d) la generación de un modelo con la matriz predictora X expandida para identificar importantes términos de productos cruzados y / o términos lineales.

E. DISEÑO DE BIBLIOTECA DE PROTEÍNA VARIANTE QUE INCORPORA LA INFORMACIÓN EVOLUTIVA

Si bien puede ser deseable variar residuos de aminoácidos en un gran número de posiciones en una sola biblioteca de variante de proteína, haciendo así que puede conducir a una biblioteca con un gran número de variantes que tienen poca o ninguna actividad, debido a combinaciones deletéreas de demasiados residuos variables. La presente

descripción proporciona una forma eficaz de la optimización de una variante de la proteína para una actividad deseada haciendo una o más bibliotecas de variante de proteína que incorporan solamente ciertas sustituciones de residuos de aminoácidos variables de un conjunto de polipéptidos parentales. El conjunto de residuos de aminoácidos variables se seleccionan para su incorporación en una biblioteca variante de la proteína basada en el contexto de la evolución de la variable de aminoácidos de residuos. Esas sustituciones que representan sustituciones evolutivamente conservadoras se incorporan en variantes de la proteína de la biblioteca.

Los cambios de aminoácidos permitidos por la evolución en general conservan pliegue y la función de las proteínas. En escalas de tiempo evolutivas relativamente cortos, permite cambios tienden a ser independientes del contexto, es decir, hacer una contribución de la aptitud "aditiva" (y trabajar bien con otros cambios). Esencialmente se puede acceder a fuentes infinitas de homólogos en cualquier escala de tiempo de divergencia deseada por cambios "permitidos" de aminoácidos para esa escala de tiempo. También existe evidencia de que las sutiles perturbaciones en la estructura de la proteína pueden tener un enorme impacto en la función (Kidokoro (1998) "Diseño de la función de las proteínas por método de perturbación física," Adv Biophys. 35: 121-143, y Shimotohno et al (2001) "La demostración de la importancia y utilidad de la manipulación de residuos de sitio no activo en el diseño de proteínas, "J Biochem (Tokio) 129: 943-948).

La presente descripción proporciona métodos para buscar espacio de secuencias al hacer sustituciones evolutivamente conservadoras para generar la diversidad con altos niveles de aptitud. De acuerdo con los métodos, por ejemplo, secuencias parentales están alineadas para determinar qué residuos varían entre secuencias parentales (es decir, son flexibles), a continuación, una matriz de sustitución evolutiva se aplica para identificar un subconjunto de los residuos variables que representan sustituciones conservativas. Una biblioteca de variantes de proteína es entonces generada que incorpora el subconjunto conservador de residuos de aminoácidos variables en las secuencias de las variantes de proteínas. Alternativamente, otras matrices de sustitución se pueden utilizar para identificar el subconjunto de residuos variables de incorporar en una biblioteca variante de la proteína. Otras matrices de sustitución adecuadas incluyen las basadas en propiedades fisicoquímicas u otros parámetros descritos en este documento. Opcionalmente, los métodos pueden ser aplicados a las secuencias individuales mediante la aplicación de un filtro o restricción definida por el usuario, tal como la cisteína, prolina y residuos de glicina permanecen sin cambios (es decir, son menos tolerantes al cambio), y luego aplicar una matriz de sustitución de los otros residuos.

Por lo general, una matriz de sustitución, tales como matrices PAM de Dayhoff (para varias distancias PAM), matrices dependientes del sitio, las matrices BLOSUM, matrices JTT, simplemente matrices binarias que capturan cualquier clasificación de aminoácidos, y similares pueden utilizarse para crear diferentes escalas de tiempo (véase, por ejemplo, Dayhoff y Eck (1968) "Un modelo de cambio evolutivo en proteínas," Atlas de la Secuencia de Proteínas y Estructura 03:33-41, y Henikoff y Henikoff (1992) "aminoácidos de sustitución de matrices de bloques de proteínas," Proc Nat'l Acad Sci EE.UU 89:10915-10919). La sintonización de la probabilidad de transición de un aminoácido a otro puede cambiar el nivel de conservación. Tanto el punto de corte de probabilidad y la propia matriz son parámetros en el modelo. Hay varias otras matrices que también están disponibles. Estas matrices pueden ser de estructura dependiente, es decir, el núcleo dentro de una proteína ha patrones de sustitución que pueden diferir de la superficie externa de la proteína, hélices puede tener diferentes patrones a partir de hebras, y similares (Koshi y Goldstein (1997) "matrices de mutación y las propiedades físico-químicas: correlaciones y consecuencias," Proteínas 27:336-344, y Koshi y Goldstein (1996) "Correlación dependiente de la estructura de matrices de mutación con discapacidad física de propiedades químicas, "Pac. Symp Biocomput. 488-499). Una matriz basada en las propiedades fisicoquímicas también se puede utilizar para seleccionar las sustituciones adecuadas. Detalles adicionales con respecto a matrices de sustitución adecuados para uso en la presente descripción se describen adicionalmente en, por ejemplo, Durbin et al., análisis de secuencias biológicas probabilísticas Modelos de proteínas y aminoácidos Cambridge University Press (1998). En el uso de cualquiera de las matrices anteriores, una biblioteca de polipéptidos de variantes que incorpora la diversidad conservadora y / o la diversidad no conservadora, se puede hacer. Para las bibliotecas no conservativas, sustituciones que son menos propensas a ocurrir bajo evolución divergente normalmente se seleccionan.

Cuando las estructuras de las proteínas de interés están disponibles, regiones / residuos pueden ser identificados que tendrán el impacto deseado en función de la proteína. Esto se puede conseguir mediante, por ejemplo, sencillo modelado de los cambios en la electrostática alrededor de sitios activos o cambios que llevan a la dinámica modificada en la proteína (Kidokoro, *supra*). La información estructural también se puede utilizar para identificar dominios / módulos que tendrán el mayor impacto y se puede limitar sus esfuerzos sólo a la región seleccionada de las proteínas.

Algoritmos de la presente descripción se pueden utilizar para construir una serie de bibliotecas, para cualquier gen dado, con un continuo de la aptitud mediana, un continuo de variación genética y fenotípica, y un alto nivel de aditivo de variabilidad genética. Los algoritmos son esencialmente "automáticos" en el sentido de que se aplican relativamente independientemente del conocimiento experto de la proteína.

Como visión general de estos métodos, la Figura 11 proporciona un gráfico que representa ciertas etapas llevadas a cabo en una realización método para buscar de manera eficiente el espacio de secuencias. Como se muestra, el

método incluye la identificación de un gen inicial o familia de genes (es decir, gen de interés) (**H1**), la obtención de secuencias de homólogos que abarcan una escala de tiempo de la evolución deseada (**H2**), y la evaluación del número y tipo de cambios de aminoácidos (por ejemplo, con respecto al polipéptido codificado por el gen inicial) que se identificó como una función del tiempo / probabilidad (P) (es decir, indicado por escala de tiempo o la probabilidad de tal mutación que se produzca en la naturaleza; nivel de conservación) (**H3**). El método también incluye la evaluación de diversidad potencial de la biblioteca como una función del tiempo / probabilidad (**H4**), y la identificación del número de posiciones variables en la escala dada de tiempo que resulta en el tamaño de la biblioteca deseada (por ejemplo, basándose en el rendimiento de la detección y de la aptitud esperada de la nueva biblioteca) (**H5**). Además, el método incluye la estimación de la aptitud mediana y la varianza de las bibliotecas como una función de la escala de tiempo de la que proviene la diversidad (**H6**), y haciendo una serie de bibliotecas que cubren la aptitud mediana y rango de varianza deseada (**H7**).

Todos estos métodos se pueden implementar para toda una alineación y / o para un conjunto de residuos usuarios específicos definidos o el uso de la información estructural para hacer bibliotecas de dominios (módulos, subdominios, etc.). Para la generación de la diversidad, estos enfoques basados en matrices pueden ser usados en conjunto con otros métodos como PCA, PLS o similares, cuyas informaciones de carga similares (por ejemplo, entropías sitio) en sitios específicos de la proteína puede atribuir importancia a las posibilidades de sustitución. Información de secuencias de consenso se puede utilizar para restringir o aumentar la diversidad en la biblioteca. Los métodos de reconstrucción de secuencia ancestrales pueden identificar de forma fiable los cambios que tuvieran lugar en el conjunto de proteínas muy temprano en el proceso evolutivo, y cambios que son de carácter adaptativo. Esto se puede utilizar de forma automática en los enfoques descritos en este documento para hacer bibliotecas deseadas.

Estos métodos incluyen típicamente diversos rigores de selección y tamaños de bibliotecas. Por ejemplo, las evaluaciones de la "fragilidad" de una proteína se realizan opcionalmente mediante estimaciones. En este cálculo, se regirán por lo general los estudios de modelo de plegamiento de proteínas (por ejemplo, ya en la literatura, etc.), los datos empíricos (por ejemplo, pantalla sobre 100-1000 visitas por biblioteca, etc.), la extrapolación de la tasa de cambios en la evolución, el tamaño de la biblioteca que se pueden cribar, y / o similares. Bibliotecas incluyen típicamente entre aproximadamente 10^3 y alrededor de 10^{12} miembros, dependiendo de los métodos de detección particular utilizados. Por ejemplo, se debe considerar la correlación de la pantalla con pantallas de mayor complejidad de aguas abajo.

Estos métodos para la búsqueda de un espacio de secuencias de alto rendimiento proporcionan muchas ventajas diferentes. En particular, el enfoque general se hace más potente y refinado a medida que los datos sobre las proteínas / pliegues de interés se acumulan. Además, el espacio de secuencia deseada se puede definir de forma automática a partir de datos filogenéticos, utilizando un ordenador. Además, la información sobre los pasos filogenéticos "seguros" (por ejemplo, sustituciones conservadoras de residuos) puede ser aprovechada para su análisis y posterior desarrollo.

En ciertos aspectos, la presente descripción proporciona un sistema para producir bibliotecas de tamaños deseados. El sistema descrito en el presente documento incluye (a) al menos un ordenador que incluye una base de datos capaz de almacenar conjuntos de cadenas de caracteres de biopolímero, y (b) del software del sistema. El software del sistema incluye una o más instrucciones lógicas para: (i) identificar uno o más homólogos de al menos una secuencia de polipéptido inicial, (ii) la comparación de las secuencias del homólogo (s) y la secuencia de polipéptido inicial; (iii) la identificación de residuos de aminoácidos variables, donde los residuos de aminoácidos variables difieren con respecto a los aminoácidos de tipo residuo de ácido en las posiciones en las secuencias de homólogo(s) y la secuencia de polipéptido inicial correspondiente; (iv) la identificación de un conjunto de residuos de aminoácidos variables conservados evolutivamente; y (v) la generación de una biblioteca de variantes de proteínas que incorpora el conjunto de residuos de aminoácidos de variables conservados evolutivamente. El software del sistema también incluye instrucciones para (iv) la identificación de posiciones de monómero variables en al menos una cadena de caracteres de biopolímero inicial de la escala de tiempo evolutivo seleccionado que dan como resultado un tamaño de la biblioteca deseada, y (v) proporcionar una serie de bibliotecas que comprenden una aptitud mediana seleccionada y el rango de varianza.

La descripción también incluye un producto de programa de ordenador para la producción de bibliotecas de tamaños deseados. El producto de programa de ordenador descrito en este documento incluye un medio legible por ordenador que tiene una o más instrucciones lógicas para: (a) identificar uno o más homólogos de al menos una cadena inicial carácter biopolímero a partir de una escala de tiempo evolutiva seleccionada, (b) el trazado de una serie de cambios de monómero para al menos una cadena de caracteres de biopolímero inicial en contra de un tiempo / probabilidad, y (c) el trazado de potencial tamaño de la biblioteca frente al tiempo / probabilidad. El producto de programa de ordenador también incluye instrucciones para (d) la identificación de posiciones de monómero variables en al menos una cadena de carácter biopolímero inicial de la escala de tiempo evolutiva seleccionada que dan como resultado un tamaño de la biblioteca deseada, y (e) proporcionar una serie de bibliotecas que comprenden una aptitud mediana seleccionada y rango de varianza.

IV. PREDICCIONES DE SECUENCIA DE ACTIVIDADES

A. USO DE REDES NEURONALES PARA IDENTIFICAR ADN O SECUENCIAS DE PROTEÍNAS CON CARACTERÍSTICAS MEJORADAS

En la presente descripción las redes neuronales se utilizan para analizar datos derivados de varios procesos de evolución artificial, incluyendo combinación aleatoria de ADN, a predecir las secuencias que han mejorado las características. En un ejemplo, este tipo de redes neuronales pueden ser utilizadas en los algoritmos genéticos para optimizar las secuencias de otras bibliotecas de variantes de proteínas. En resumen, los métodos incluyen el uso de los datos de cada ronda de, por ejemplo, un procedimiento de barajado como un conjunto de entrenamiento de una red neuronal. Una vez que una red neural ha sido entrenado, secuencias de cadenas de caracteres se pueden "ensayar" *in silico* usando la red entrenada. Las secuencias que la red identifica que tienen características mejoradas después se añaden normalmente a las siguientes rondas de barajado, o se sintetizan *de novo*. Los sistemas utilizados para evaluar estas secuencias de cadenas de caracteres después se anotan tienen en cuenta opcionalmente no sólo la evaluación predicha de red neuronal, sino también una puntuación de cuántas secuencias de cadenas de caracteres derivadas (por ejemplo, variantes de cadena de caracteres de de las secuencias de cadenas de caracteres nuevamente predichas) también tienen un alto puntaje de la red neuronal. Por ejemplo, si la secuencia de cadena de caracteres A se mutó en 1000 variantes de cadena de caracteres, y cada variante se puntuó de acuerdo a la red, el porcentaje de las variantes de cadena de caracteres que la puntuación por encima de cierto punto de corte en la red neuronal se cuentan opcionalmente. Además, estos datos se pueden combinar con la puntuación de redes neuronales de la secuencia de cadena de caracteres A para producir una puntuación final. Tal puntuación representaría no sólo lo que predijo la red para esa secuencia, sino también la forma en que la secuencia es probable que se mutaría en secuencias igual de buenas o mejores.

Para ilustrar aún más, la Figura 13 proporciona un diagrama que muestra ciertos pasos llevados a cabo en una realización de un método de predicción de series de caracteres que incluyen las propiedades deseadas. Como se muestra, los métodos incluyen la evolución de al menos una cadena de caracteres de los padres (por ejemplo, una pluralidad de cadenas de caracteres de los padres, etc.) usando al menos un procedimiento de evolución artificial para producir al menos una población de cadenas de caracteres evolucionados artificialmente (**II**). Procedimientos de evolución artificial llevados a cabo en las cadenas de caracteres normalmente se realizan de manera reiterada para producir múltiples poblaciones de cadenas de caracteres evolucionadas artificialmente, que varias poblaciones de cadenas de caracteres evolucionadas artificialmente se utilizan para entrenar la red neuronal. Los métodos también incluyen la selección o cribado de la población de cadenas de caracteres evolucionados artificialmente para al menos una propiedad deseada (por ejemplo, una propiedad física, una propiedad catalítica, o similar, que es propiedad mejorada con respecto a la cadena de caracteres parentales) para producir una población de cadenas de caracteres de evolución artificial seleccionados (**12**). Los métodos también incluyen entrenamiento de una red neuronal con la población de cadenas de caracteres de evolución artificial seleccionados para producir una red neuronal entrenada (**13**). A partir de entonces, los métodos incluyen cadenas de caracteres que predicen que incluyen, o es probable que incluyan, la propiedad deseada usando la red neuronal entrenada (**14**). Detalles adicionales relativas a las redes neuronales son proporcionadas anteriormente.

En ciertas realizaciones, los métodos incluyen además la repetición de los pasos **11** y **12** utilizando la población de cadenas de caracteres de evolución artificial seleccionados en el paso **12** como al menos una cadena de caracteres parental en un paso repetido **11**. En estas realizaciones, los métodos opcionalmente incluyen además el uso de la población de cadenas de caracteres de evolución artificial seleccionados de al menos un paso repetido **12** para entrenar aún más la red neural en el paso **13**. Las cadenas de caracteres parentales normalmente corresponden a polinucleótidos o polipéptidos. En algunas realizaciones, los métodos opcionalmente incluyen además la síntesis de polinucleótidos o polipéptidos que corresponden a las cadenas de caracteres previstos en el paso **14**. En otras realizaciones, los métodos incluyen además la repitición de las etapas **11-14**, utilizando al menos una de las cadenas de caracteres previstos en el paso **14** como cadena de caracteres parental en un paso repetido **14**. Típicamente, los métodos incluyen además el uso de la red neuronal entrenada como filtro para inclinar la producción biblioteca hacia miembros de la biblioteca activos.

En particular, el paso **14** incluye típicamente la puntuación de varias cadenas de caracteres utilizando un sistema de puntuación de la red neuronal entrenada para predecir las cadenas de caracteres con la propiedad deseada. El sistema de puntuación se ubica generalmente cadenas de caracteres puntuadas. Además, el sistema de puntuación normalmente representa un número de cadenas de caracteres de progenie de cada cadena de caracteres que incluye una puntuación por encima de una puntuación seleccionada. Por ejemplo, el número de cadenas de caracteres de la progenie típicamente incluyen, por ejemplo, entre aproximadamente dos y aproximadamente 10^5 cadenas de caracteres de progenie. En general, el sistema de puntuación combina la puntuación de cada cadena de caracteres con la puntuación de cada cadena de caracteres correspondientes de progenie para producir una puntuación final. La puntuación final proporciona una medida de la probabilidad de las cadenas de caracteres mutantes en cadenas de caracteres de la progenie que se mejoran con relación a las cadenas de caracteres.

Los procedimientos de evolución artificial utilizados en el paso **11** opcionalmente se llevan a cabo *in silico* y, en consecuencia, por lo general incluyen la aplicación de los operadores genéticos para cadenas de caracteres

parentales para producir la población de cadenas de carácter de evolución artificial. Operadores genéticos ejemplares opcionalmente utilizados en estos métodos incluyen, por ejemplo, una mutación de al menos una cadena de caracteres parentales o subcadenas de al menos una cadena de caracteres parentales, una multiplicación de al menos una cadena de caracteres parentales o subcadenas de al menos una cadena de caracteres parentales, una fragmentación de al menos una cadena de caracteres parentales en subcadenas, un cruce de cadenas de caracteres parentales o subcadenas de las cadenas de caracteres parentales, una ligadura de cadenas parentales de caracteres o subcadenas de las cadenas de caracteres parentales, un cálculo de elitismo, un cálculo de la homología de secuencia o similitud de secuencia de una alineación que comprende cadenas de caracteres de los padres, un uso recursivo de al menos uno de los uno o más operadores genéticos, una aplicación de un operador de aleatoriedad a al menos una cadena de caracteres parentales o subcadenas de la al menos una cadena de caracteres parentales, una mutación de deleción de una o más cadenas de caracteres parentales o subcadenas de la o cadenas de caracteres parentales, una mutación de inserción en al menos una cadena de caracteres parentales o subcadenas de la cadena de caracteres parentales, una resta de cadenas de caracteres parentales con secuencias inactivas, una selección de cadenas de caracteres parentales con secuencias activas, una muerte de cadenas de caracteres parentales o subcadenas de las cadenas de caracteres parentales, o similares.

La divulgación también proporciona un sistema de ordenador para la predicción de las cadenas de caracteres que incluyen las propiedades deseadas. El sistema incluye (a) un sistema de ordenador que incluye una red neuronal y una base de datos capaz de almacenar cadenas de caracteres, y (b) software del sistema. El software del sistema incluye una o más instrucciones lógicas para (i) cambiar al menos una cadena de caracteres parentales utilizando al menos un procedimiento de evolución artificial para producir al menos una población de cadenas de caracteres de evolución artificial, y (ii) la selección o cribado de la población de manera de cadenas de caracteres de evolución artificial de al menos una propiedad deseada para producir una población de cadenas de caracteres de evolución artificial seleccionados. El software también incluye instrucciones para (iii) el entrenamiento de la red neuronal con la población de cadenas de caracteres de evolución artificial seleccionados para producir una red neuronal entrenada, y (iv) la predicción de una o más cadenas de caracteres que componen por lo menos una propiedad deseada mediante la red neural entrenada.

En otro aspecto, la descripción se refiere a un producto de programa de ordenador para la predicción de las cadenas de caracteres que incluyen las propiedades deseadas. El producto de programa de ordenador incluye un medio legible por ordenador que tiene una o más instrucciones lógicas para (a) la evolución de al menos una cadena de caracteres parental usando al menos un procedimiento de evolución artificial para producir al menos una población de cadenas de caracteres de evolución artificial, y (b) selección o cribado de la población de cadenas de caracteres de evolución artificial para al menos una propiedad deseada para producir una población de cadenas de caracteres de evolución artificial seleccionados. El producto también incluye instrucciones para (c) la formación de una red neuronal con la población de cadenas de carácter de evolución artificial seleccionados para producir una red neuronal entrenada, y (d) la predicción de una o más cadenas de caracteres que componen al menos una propiedad deseada usando la red neuronal entrenada. Sistemas y software se describen con más detalle en el presente documento.

B. USO DE ALGORITMOS DE PATRÓN O MOTIVO DE BÚSQUEDA PARA ANALIZAR EL ESPACIO CON SECUENCIA

Hay muchos programas informáticos disponibles para la búsqueda y hallazgo y motivos dentro de un grupo de secuencias. Normalmente, estos programas se limitan a la caracterización de secuencias como parte de una familia amplia de proteínas o no. En la presente invención, los programas de búsqueda de motivos se utilizan para caracterizar y predecir la actividad de proteínas, p.ej. proteínas de evolución artificial. Por ejemplo, las secuencias positivas (por ejemplo, los que tienen un nivel deseado de aptitud), secuencias negativas (por ejemplo, los que carecen de un nivel deseado de aptitud), y los parentales se introducen opcionalmente en los programas de búsqueda de patrón por separado. Sin embargo, todos los tipos de secuencias se introducen opcionalmente en el programa de patrón de conclusión, junto, por ejemplo, para aumentar la sensibilidad a la búsqueda de cualesquiera patrones. Debido a la alta homología general de las secuencias positivas, programas de búsqueda de motivos suelen encontrar muchos motivos o patrones que existen dentro de cada grupo de secuencias. Los patrones se obtuvieron opcionalmente de acuerdo con una frecuencia de ocurrencia en cada grupo, a una frecuencia de ausencia de cada grupo de secuencias, y / o similares. Además, los patrones detectados también se introducen opcionalmente en otro algoritmo de reconocimiento de patrones tal como una red neural. Una vez que el reconocimiento de patrones y la puntuación son completos, secuencias hipotéticas se califican con el fin de encontrar las secuencias adicionales que serán o son más propensos a tener la actividad / propiedad deseada. Además, el análisis PCA se lleva a cabo opcionalmente en el patrón de resultados de la búsqueda para determinar si hay combinaciones de motivos o patrones que son predictivos de la actividad, que luego se utilizan para anotar las secuencias de proteínas adicionales. Estos métodos normalmente se implementan en realizaciones web u otras basadas en software, y opcionalmente se acoplan con herramientas de análisis bioinformáticos adicionales, tales como el análisis cruzado, arrastrando los análisis, la creación de oligo, análisis estructural, etc. con el fin de vender los kits de biología molecular para barajada, venta de oligonucleótidos, u otros software o servicios bioinformáticos.

En ciertas realizaciones, los árboles de búsqueda se generan, que se basan, por ejemplo, en un método de

puntuación para la organización de los patrones, o grupos de patrones de tal manera que permita recorrer el árbol en lugar de tratar todos los patrones posibles, y la combinación de patrones. Por ejemplo, los patrones están opcionalmente marcados por la frecuencia con que aparecen en las secuencias positivas / negativas. En lugar de los patrones individuales, análisis de PCA o similares se lleva a cabo opcionalmente para determinar las combinaciones de patrones para cada uno de los nodos. Para ilustrar, los resultados de los patrones de búsqueda de las secuencias positivas y negativas se analizan utilizando opcionalmente PCA. Un valor de corte de carga se utiliza típicamente para cada componente principal y patrones resultantes (por ejemplo, una lista de patrones) corresponderían entonces a los nodos del árbol.

Además, los patrones se puntúan opcionalmente con un valor que se refiere, por ejemplo, al contenido relativo de información, importancia, aptitud etc., así como un valor de la actividad prevista. Estos se utilizan opcionalmente de nuevo para entrenar las redes neuronales o para construir un árbol de decisiones para clasificar o puntuar proteínas hipotéticas u otros biopolímeros. Por ejemplo, si se demuestra que el patrón AAA.GAW es el más importante, entonces proteínas hipotéticas son normalmente revisadas en función de si tienen el siguiente patrón más importante en esa subrama. Este proceso se continúa opcionalmente con el siguiente patrón más importante dado, por ejemplo, que el primero se encontró o no se encontró y se clasifica la secuencia basada en esa secuencia. El "contiene" y "no contiene" sub-árboles pueden incluir nodos similares (es decir, patrones), o pueden no depender de la importancia que se da a un patrón particular, su linaje de nodo parental. Para ilustrar aún más, la Figura 14 muestra esquemáticamente un ejemplo de árbol de organización. En el ejemplo, si un patrón tiene los tres patrones AAA.GAW, AAA.G.W.W, y GPPW, entonces su probabilidad de tener la actividad deseada es de 60%. Además, se podría basar en el hecho de que 60% de las secuencias positivas tienen estos tres patrones.

La Figura 15 es un gráfico que representa ciertas etapas llevadas a cabo en una realización de los métodos de la predicción de las propiedades de las cadenas de polipéptido de objetu=ivo de caracteres (por ejemplo, al menos una cadena de caracteres hipotéticos de polipéptido, etc.). Como se muestra, los métodos incluyen la identificación de uno o más motivos comunes a dos o más miembros de una población de variantes de cadena de caracteres de polipéptidos en el que al menos un subconjunto de la población de variantes de cadena de caracteres de polipéptidos incluye al menos una propiedad (por ejemplo, una propiedad funcional, propia de la estructura, y / o similares), para producir un conjunto de datos con motivos (**J1**). En ciertas realizaciones, una familia filogenética incluye las variantes de cadenas de caracteres de polipéptidos. Al menos uno de los uno o más motivos típicamente incluyen una o más subcadenas de caracteres. Típicamente, al menos un polipéptido de objetivo incluye una población de cadenas de caracteres de objetivo de polipéptido. En estas realizaciones, la población de cadenas de caracteres de objetivo de polipéptido se produce generalmente por uno o más procedimientos de evolución artificial. Los métodos también incluyen **J2** la correlación de al menos un motivo partir de los datos con motivos establecidos con la al menos una propiedad para producir una función de puntuación motivo, y **J3** anotando la cadena de caracteres polipéptido al menos un objetivo utilizando la función de puntuación motivo para predecir la al menos una propiedad de la al menos un objetivo cadena de caracteres polipéptido. Al menos un paso de estos métodos se lleva a cabo típicamente en un sistema digital o basado en la web. Opcionalmente, los métodos incluyen además la síntesis de un polipéptido correspondiente a la cadena de caracteres de objetivo de polipéptido. Una opción adicional incluye someter el polipéptido, o un polinucleótido que codifica el polipéptido, uno o más procedimientos de evolución artificiales.

Funciones de puntuación de motivo se producen utilizando técnicas de variación. Por ejemplo, el paso **J2** incluye opcionalmente la puntuación de los motivos o combinaciones de los motivos según las frecuencias de ocurrencia de variantes de cadenas de variantes de caracteres positivos de polipéptido o de polipéptido de cadena de caracteres negativos para producir la función de puntuación de motivo. En algunas realizaciones, el paso **J2** incluye la puntuación de los motivos, o combinaciones de los motivos, con un valor relativo al contenido de información relativa y / o aptitud relativa. En otras formas de realización, el paso **J2** incluye la puntuación de los motivos, o combinaciones de los motivos, con valores relacionados con la actividad de predicción relativa. En todavía otras formas de realización, el paso **J2** incluye la determinación de un número de veces que uno o más motivos ocurren en o están ausentes de los dos o más miembros de la población de variantes de cadenas de caracteres de polipéptido.

En ciertas realizaciones, la población de variantes de cadena de caracteres de polipéptido incluye uno o más grupos de variantes de cadena de caracteres de polipéptido. Cada grupo de variantes cadena de caracteres polipéptido incluye opcionalmente, por ejemplo, de carácter positivo polipéptido variantes de polipéptidos de cadena negativa, variantes de cadenas de caracteres y / o variantes de cadena de caracteres de polipéptido parentales. Las variantes de cadena de caracteres de polipéptido se producen típicamente por o corresponden a polipéptidos producidos por uno o más procedimientos de evolución artificial. Al menos uno (y por lo general más de uno) paso de las una o más técnicas de evolución artificial se realiza generalmente *in silico*.

En formas de realización preferidas, al menos el paso **J1** se lleva a cabo en al menos un dispositivo de lógica que incluye al menos un primer algoritmo de reconocimiento de motivo, cuyo primer algoritmo de reconocimiento de motivo identifica uno o más motivos. Típicamente, cada etapa del procedimiento se realiza en al menos un dispositivo lógico. Opcionalmente, los métodos incluyen además la producción de al menos una clasificación libre (por ejemplo, al menos un árbol de clasificación y regresión (CART), etc.) para organizar los motivos del conjunto de

datos de motivo. Por ejemplo, el al menos un árbol de clasificación normalmente permite buscar los datos con motivos establecidos sin tratar todos los motivos o combinaciones de motivos en el conjunto de datos con motivos.

En algunas realizaciones, los métodos incluyen además la realización de análisis de componentes principales en el conjunto de datos con motivos para identificar una o más combinaciones de motivos que son predictivos de la al menos una propiedad deseada. Opcionalmente, los métodos incluyen además la realización de un análisis parcial de mínimos cuadrados de los datos con motivos establecidos para identificar una o más combinaciones de motivos que son predictivos de la propiedad deseada. Una o más combinaciones de motivos identificadas se utilizan normalmente para perfeccionar la función de puntuación de motivo. Además, los métodos opcionalmente incluyen además la producción de al menos una clasificación libre (por ejemplo, al menos una clasificación y regresión árbol, etc.) para organizar una o más combinaciones de motivos. En estas realizaciones, una o más combinaciones de motivos incluyen típicamente nodos en al menos un árbol de clasificación. Por lo general, los permisos de al menos un árbol de clasificación que buscan los datos con motivos se establecen sin tratar todos los motivos o combinaciones de motivos en el conjunto de datos con motivos. En ciertas otras realizaciones, los métodos incluyen además el sometimiento de los datos con motivos establecidos a por lo menos un segundo algoritmo de reconocimiento de patrones, cuyo segundo algoritmo de reconocimiento de patrones identifica al menos un motivo adicional comunes al menos a dos miembros de la población de variantes de cadenas de caracteres de polipéptido. Por ejemplo, el segundo reconocimiento de patrones de algoritmo incluye opcionalmente una red neural. Las redes neuronales se describen con más detalle en el presente documento.

La divulgación también proporciona un sistema para predecir al menos una propiedad de al menos un objetivo cadena de caracteres de polipéptido. El sistema incluye (a) al menos un ordenador que incluye una base de datos capaz de almacenar cadenas de caracteres, y (b) el software del sistema. El software del sistema incluye una o más instrucciones lógicas para (i) la identificación de uno o más motivos comunes a dos o más miembros de una población de variantes de cadena de caracteres de polipéptido, en el que al menos un subconjunto de la población de variantes de cadena de caracteres de polipéptido comprende al menos una propiedad, para producir un conjunto de datos con motivos. El software también incluye instrucciones para (ii) la correlación de al menos un motivo a partir de los datos con motivos establecidos con al menos una propiedad para producir una función de puntuación de motivo, y (iii) anotando la cadena de caracteres de polipéptido al menos un objetivo utilizando la función de puntuación de motivo para predecir al menos una propiedad de al menos una cadena de caracteres de polipéptido de objetivo.

Además, la descripción también se refiere a un producto de programa de ordenador para la predicción de al menos una propiedad de al menos una cadena de caracteres de polipéptido de objetivo. El producto de programa de ordenador incluye un medio legible por ordenador que tiene una o más instrucciones lógicas para (a) la identificación de uno o más motivos comunes a dos o más miembros de una población de variantes de cadenas de caracteres de polipéptido, en el que al menos un subconjunto de la población de variantes de cadenas de caracteres de polipéptido comprende el establecimiento al menos una propiedad, para producir un conjunto de datos de motivo. El producto de programa de ordenador también incluye instrucciones para (b) correlacionar al menos un motivo a partir de los datos con motivos establecidos para la al menos una propiedad para producir una función de puntuación de motivo, y (c) puntuación de al menos una cadena de caracteres de polipéptido de objetivo usando la función de puntuación motivo para predecir al menos una propiedad de al menos una cadena de caracteres de polipéptido de objetivo.

C. EVOLUCIÓN DIRIGIDA IN SILICO CON CHEQUEO FUNCIONAL MEDIANTE PCA Y REDES NEURONALES

En ciertas realizaciones, al menos un miembro del conjunto de cadenas de caracteres parentales se obtiene a partir de al menos una base de datos. En algunas de estas realizaciones, al menos un miembro incluye sustancialmente todas las cadenas de caracteres disponibles de la base de datos. Típicamente, al menos un miembro del conjunto de cadenas de caracteres parentales es producido por, o corresponde al menos con un polinucleótido o al menos un polipéptido producido por uno o más procedimientos de evolución artificial. Al menos un paso de los procedimientos de evolución artificial se realiza típicamente *in silico*. En algunas realizaciones, el conjunto de cadenas de caracteres parentales corresponde a un conjunto de polinucleótidos o polipéptidos parentales.

La divulgación también proporciona un sistema para la asignación de una actividad a una cadena de caracteres. El sistema incluye (a) al menos un ordenador que incluye una base de datos capaz de almacenar cadenas de caracteres, y (b) el software del sistema. El software del sistema incluye una o más instrucciones lógicas para (i) la selección de un conjunto de cadenas de caracteres parentales por lo menos una actividad para producir un conjunto de cadenas de caracteres parentales seleccionados, y (ii) el sometimeinto del conjunto de cadenas de caracteres parenatales seleccionados para uno o más procedimientos de evolución artificiales para producir un conjunto de cadenas de caracteres evolucionados. El software del sistema también incluye instrucciones para (iii) seleccionar el conjunto de cadenas de caracteres evolucionados para al menos una actividad para producir un conjunto de cadenas de caracteres evolucionados seleccionados, (iv) proporcionar un gráfico de sucesión en la actividad para el conjunto de variantes de cadena de caracteres y (v) la predicción de por lo menos una actividad de una o más cadenas de caracteres del gráfico de sucesión en la actividad.

Además, la descripción proporciona un producto de programa de ordenador para la predicción de las actividades de

cadena de caracteres. El producto de programa de ordenador incluye un medio legible por ordenador que tiene una o más instrucciones lógicas para (a) seleccionar un conjunto de cadenas de caracteres parentales para por lo menos una actividad para producir un conjunto de cadenas de caracteres parenteles seleccionados, y (b) el sometimiento del conjunto de cadenas de caracteres parentales seleccionados a uno o más procedimientos de evolución artificial para producir un conjunto de cadenas de caracteres evolucionados. El producto también incluye instrucciones para (c) seleccionar el conjunto de cadenas de caracteres evolucionados para al menos una actividad para producir un conjunto de cadenas de caracteres evolucionados seleccionados, (d) proporcionar un gráfico de sucesión-actividad para el conjunto de variantes de cadena de carácter, y (e) predecir al menos una actividad de una o más cadenas de caracteres de de la trama de actividad de secuencia.

V. TÉCNICAS EXPERIMENTALES

A. BIBLIOTECAS DE VARIANTE DE PROTEÍNA

Bibliotecas de variantes de proteínas se pueden generar utilizando cualquiera de una variedad de métodos que son bien conocidos por expertos en la técnica. Estas bibliotecas se preparan típicamente por la expresión, ya sea in vivo o in vitro, de una biblioteca de polinucleótidos diversos. Bibliotecas de polinucleótidos diversos pueden generarse por aplicación de un "procedimiento de generación de diversidad" a uno o más polinucleótidos "parentales".

Tal como se utiliza aquí, el término " procedimiento de generación de diversidad" se refiere a un método que modifica la secuencia de un polinucleótido parental, y de forma concomitante el polipéptido que codifica, generando de ese modo una biblioteca de variantes de polinucleótidos que difieren entre sí con respecto a la secuencia. Procedimientos de generación de diversidad que son adecuados para su uso en la práctica de la presente invención incluyen tanto mutagénesis como métodos basados en la recombinación, o una combinación de ambos. La expresión de la biblioteca de variante de polinucleótido resultante genera una biblioteca de variantes de polipéptido

Bibliotecas de variante de proteína empleados en la práctica de la presente invención se pueden fabricar de una manera "a ciegas", donde las moléculas son variantes de la proteínas generadas sin conocimiento previo de sus secuencias de aminoácidos (es decir, donde las secuencias de variantes de polinucleótidos no se conocen antes de la expresión en una biblioteca de variante de proteína). Alternativamente, las secuencias de aminoácidos que codifica variantes de proteína pueden diseñarse a priori, seguido por el paso de hacer realidad las moléculas físicas utilizando métodos conocidos por los expertos en la técnica. Estos métodos incluyen la expresión de polinucleótidos generados por, por ejemplo, síntesis de genes a través de la ligadura y / o el conjunto de oligonucleótidos mediados por polimerasa y mutagénesis de un polinucleótido parental, usando métodos conocidos en la técnica. Los métodos adecuados para el diseño de secuencias de aminoácidos de variantes de proteínas sistemáticamente variadas incluyen el diseño de métodos de experimento (DOE), que se describen en más detalle en este documento.

La mutagénesis de polinucleótido es un método adecuado para la generación de las variantes de proteínas empleados en la práctica de la presente invención. Tales métodos incluyen, por ejemplo, la reacción en cadena de polimerasa propensa a errores (PCR), la mutagénesis específica de sitio de casetes de mutagénesis in vivo, en métodos de mutagénesis, y similares. En la PCR propensa a errores, la PCR se realiza en condiciones en que la fidelidad de copia de la polimerasa de ADN es baja, de manera que se obtiene una alta tasa de mutaciones puntuales a lo largo de toda la longitud del producto de PCR. Véase, por ejemplo, Leung et al. (1989) Technique 1: 11-15 y Caldwell et al. Métodos PCR Aplic (1992) 2:28-33. mutaciones específicas del sitio pueden introducirse en una secuencia de polinucleótidos de interés usando mutagénesis dirigida a oligonucleótidos. Ver Reidhaar-Olson et al. (1988) Science 241: 53-57. Del mismo modo, la mutagénesis de cassette se puede utilizar en un proceso que sustituye a una pequeña región de una molécula de ADN de doble cadena con un casete de oligonucleótido sintético que difiere de la secuencia nativa. La mutagénesis in vivo se puede utilizar para generar mutaciones aleatorias en cualquier ADN clonado de interés mediante la propagación de la ADN en una cepa de célula huésped propensa a las mutaciones que generan, por ejemplo, en una cepa de *E. coli* que lleva mutaciones en uno o más de las vías de reparación del ADN. Estas cepas "mutantes" tienen una mayor tasa de mutación al azar que la de uno de los padres de tipo silvestre. La propagación del ADN en una de estas cepas genera finalmente mutaciones aleatorias dentro del ADN. Los métodos de mutagénesis son generalmente bien conocidos por los expertos en la técnica y se describen ampliamente en otra parte. Véase, por ejemplo, Kramer et al. (1984) Cell 38: 879-887; Carter et al. (1985) Nucl. Acids. Res. 13: 4431-4443; Carter (1987) Methods in Enzymol 154: 382-403; Eghtedarzadeh y Henikoff (1986) Nucl. Acids. Res. 14: 5115; Wells et al. 30 (1986) Phil Trans. R Soc. Lond A 317: 415-423; Nambiar et al. (1984), Science 223: 1299-1301; Sakamar y Khorana (1988) Nucl. Acids. Res. 14: 6361-6372; Wells et al. (1985) Gene 34: 315-323; Grundström et al. (1985) Nucl. Acids. Res. 13: 3.305 a 3316; Mandecki (1986) Proc. Natl. Acad. Sci. USA. 83: 7.177-7.181; Arnold (1993) Current Opinion in Biotechnology 4: 450-455); Anal Biochem 254(2): 157-178; Dale et al. (1996) Methods Mol Biol. 57: 369-374; Smith (1985) Ann Rev. Genet 19: 423-462; Botstein y Shortle (1985) Science 229: 1193-1201; Carter (1986) Biochem J 237: 1-7; Kunkel (1987) en Nucleic Acids & Molecular Biology Eckstein, F. y Lilley, D.M.J. eds, Springer Verlag, Berlín.; Kunkel (1985) Proc Natl. Acad. 5 Sci EE.UU. 82: 488-492; Kunkel et al. (1987) Methods in Enzymol 154, 367-382; y Bass et al. (1988) Science 242: 240-245; Methods in Enzymol 100: 468-500 (1983); Métodos en Enzymot 154: 329-350 (1987); Zoller y Smith (1982) Nucleic Acids Res. 10: 6487 a 6.500; Zolter y Smith (1983) Methods in Enzymol 100: 468-500; y Zotler y Smith (1987) Methods in Enzymol 154: 329-350); Taylor et al. (1985) NUCT 10 Acids Res. 13: 8749 hasta 8764; Taylor et al. (1985) Nucl Acids Res. 13: 8765 hasta

8787 (1985); Nakamaye y Eckstein (1986) nuci Acids Res. 14: 9679-9698; Sayers et al. (1988) NUCT Acids Res. 16: 791-802; Sayers et al. (1988) Nucl Acids Res. 16: 803814); Kramer et al. (1984) Nucl Acids Res. 12: 9441 hasta 9.456; Kramer y Fritz (1987) Methods in Enzymol 154: 350-367; Kramer et al. (1988) Nucl Acids Res. 16: 7207; 15 y Fritz et al. (1988) Nucl Acids Res. 16: 6.987 hasta 6.999.

Equipos para la mutagénesis, la construcción de la biblioteca y otros métodos de generación de diversidad están disponibles comercialmente. Por ejemplo, los equipos están disponibles de, por ejemplo, Stratagene (por ejemplo, QuickChange™ equipo de mutagénesis dirigida al sitio, y Chameleon™, equipo de mutagénesis de doble cadena dirigida al sitio), Bio / CAN Scientific, Bio-Rad (por ejemplo, utilizando el método Kunkel mencionado anteriormente), Boehringer Mannheim Corp., Clonetech Laboratories, Tecnologías de ADN, Epicentro Technologies (por ejemplo, 5 Prime 3 equipo Prime); Genpak inc, Lemargo inc, Life Technologies (Gibco BRL), New England Biolabs, Pharmacia Biotech, Promega Corp., Quantum Biotechnotogies, Amersham International plc (por ejemplo, usando el método de Eckstein al que se refiere más arriba), y Anglian Biotechnology Ltd. (por ejemplo, utilizando el método de Carter / Winter al que se refiere anteriormente).

Métodos basados en la recombinación también son adecuados para generar una biblioteca diversa de variantes de polinucleótidos que se pueden expresar para generar una biblioteca variante de la proteína. Estos métodos también se conocen como barajado de ADN. En estos métodos, los polinucleótidos se recombinan, ya sea in vitro o in vivo, para generar una biblioteca de variantes de polinucleótidos. En los métodos basados en la recombinación, fragmentos de ADN, los amplicones de PCR, y / o oligonucteotides sintéticos que corresponden colectivamente en secuencia a algunos o la totalidad de la secuencia de uno o más polinucleótidos parentales se recombinan para generar una biblioteca de variantes de polinucleótidos de la polinucleótido parental(es). El proceso de recombinación puede ser mediado por la hibridación de los fragmentos de ADN, amplicones de PCR, y / o oligonucleótidos sintéticos entre sí (por ejemplo, como dúplex se superponen parcialmente), o a un pedazo más grande de ADN, tal como una plantilla de longitud completa. Dependiendo del formato de recombinación empleado, ligasa y / o de la polimerasa se puede utilizar para facilitar la construcción de un polinucleótido de longitud completa. PCR ciclismo se utiliza típicamente en formatos que emplean solamente una polimerasa. Estos métodos son generalmente conocidos por los expertos en la técnica y se describen extensamente en otro lugar. Véase, por ejemplo, Soong, N. et al. (2000) Genet 25 (4): 436-439; Stemmer, et al. (1999) Tumor Targeting 4: 1-4; Ness et al. (1999) Nature Biotechnology 17: 893-896; Chang et al. (1999) Nature Biotechnology 10 17: 793-797; Minshull y Stemmer (1999) Current Opinion in Chemical Biology 3: 284-290; Cristianos et al. (1999) Nature Biotechnology 17: 259-264; Crameri et al. (1998) Nature 391: 288-291; Crameri et al. (1997) Nature Biotechnology 15: 436-438; Zhang et al. (1997) Proc Natl. Acad. Sci EE.UU. 94: 4504-4509; Patten et al. (1997) Current Opinion in Biotechnology 8: 724-733; Crameri et al. (1996) Nature Medicine 2 15: 100-103; Crameri et al. (1996) Nature Biotechnology 14: 315-319; Gates, et al. (1996) Revista de Biología Molecular 255: 373-386; Stemmer (1996) En: La Enciclopedia de la Biología Molecular. VCH Publishers, Nueva York. ppA4'7-45'7; Crameri y Stemmer (1995) 18: 194-195 BioTechnigues; Stemmer et al, (1995) Gen 164:. 49-53; Stemmer (1995) "La evolución de la Computación Molecular" 20 Science 270: 1510; Stemmer (1995) Bio / Technology 13: 549-553; Stemmer (1994) Nature 370: 389-391; y Stemmer (1994) Proc Natl. Acad. Sci EE.UU. 91: 1074710751; Dador y Arnold (1998) Current Opinion in Chemical Biology 2: 335-338; Zhao et al. (1998) Nature Biotechnology 16: 258-261; Coco et al. (2001) Nature Biotechnology 19: 354-359; La patente de EE.UU.. Nos. 5.605.793, 5.811.238, 5.830.721, 25 5.834.252, 5.837.458, WO 95/22625, WO 96/33207, WO 97/20078, WO 97/35966, WO 99/41402, WO 99/41383, WO 99/41369, WO 99 / 41368, WO 99/23107, WO 99/21979, WO 98/31837, WO 98/27230, WO 98/27230, WO 00/00632, WO 00/09679, WO 98/42832, WO 99/29902, WO 98 / 41653, WO 98/41622, WO 98/42727 y WO 00/18906, WO 00/04190, WO 00/42561, WO 00/42559, WO 00/42560 30, WO 01/23401, WO 00/20573, WO 01/29211, WO 00/46344, y WO 01/29212.

Polinucleótidos parentales empleados en los procesos de recombinación a los que se hace referencia anteriormente pueden ser cualquiera de los polinucleótidos de tipo silvestre o polinucleótidos de origen no natural. En una realización de la presente descripción, las variantes de proteína que tienen secuencias sistemáticamente variadas se preparan por recombinación de dos o más polinucleótidos parentales, seguido de expresión. En algunas realizaciones, los polinucleótidos parentales son miembros de una misma familia de genes. Tal como se utiliza aquí, el término "familia de genes" se refiere a un conjunto de genes que codifican polipéptidos que exhiben el mismo tipo, aunque no necesariamente en el mismo grado de una actividad.

Ácidos polinucleicos pueden recombinarse in vitro por cualquiera de una variedad de técnicas, incluyendo por ejemplo, la digestión con ADNasa de ácidos nucleicos a ser recombinados seguido de la ligadura y / o el reensamblaje por PCR de los ácidos nucleicos. Por ejemplo, mutagénesis por PCR sexual se puede utilizar en la que fragmentación al azar (o pseudoaleatoria, o incluso no aleatoria) de la molécula de ADN es seguido por recombinación, en base a la similitud de secuencia, entre las moléculas de ADN con secuencias diferentes pero relacionadas de ADN, in vitro, seguido de la fijación de la cruce por extensión en una reacción en cadena de la polimerasa. Este proceso, y muchas variantes de procedimiento se describe, por ejemplo, en Stemmer (1994) Proc Natl. Acad. Sci EE.UU. 91: 10747 a 10751.

Métodos de recombinación sintéticos también se pueden utilizar, en la que oligonucleótidos correspondientes a los objetivos de interés se sintetizan químicamente y vuelven a montar en la PCR o la ligadura de las reacciones que incluyen oligonucleótidos que corresponden a más de un polinucleótido parental, generando así nuevos

polinucleótidos recombinados. Los oligonucleótidos pueden hacerse por métodos de adición estándares de nucleótidos, o se pueden hacer, por ejemplo, mediante procedimientos sintéticos tri-nucleótidos. Los detalles relacionados con estos enfoques se encuentran en las referencias mencionadas anteriormente, por ejemplo, el documento WO 00/42561 por Crameri et al. "Olgonucleótido mediado por ácido nucleico de recombinación;" El documento WO 01/23401 por Welch et al., "El uso de codón variado de síntesis de oligonucleótidos para barajada sintética;" WO 00/42560 por Selifonov et al, "Los métodos para elaborar cadenas de caracteres, los polinucleótidos y polipéptidos que tienen las características deseadas;" WO 00/42559 y por Selifonov y Stemmer "Métodos de estructuras de inserción de datos para uso en simulaciones evolutivas."

Los polinucleótidos también pueden recombinarse in vivo, por ejemplo, al permitir ls recombinación que se produzca entre los ácidos nucleicos en las células. Muchos de tales formatos in vivo de recombinación se exponen en las referencias mencionadas anteriormente. Tales formatos proporcionan opcionalmente la recombinación directa entre los ácidos nucleicos de interés, o proporcionan la recombinación entre los vectores, virus, plásmidos, etc., que comprenden los ácidos nucleicos de interés, así como otros formatos. Los detalles relativos a tales procedimientos se encuentran en las referencias citadas en este documento.

Muchos métodos para acceder a la diversidad natural, por ejemplo, por hibridación de diversos ácidos nucleicos o fragmentos de ácido nucleico a las plantillas de cadena sencilla, seguido de polimerización y / o la ligadura para regenerar secuencias de longitud completa, seguido opcionalmente por la degradación de las plantillas y la recuperación de los resultantes ácidos nucleicos modificados se puede utilizar de manera similar. Estos métodos se pueden utilizar en sistemas físicos o se pueden realizar en los sistemas informáticos de acuerdo con realizaciones específicas de la divulgación. En un método que emplea un molde de cadena sencilla, la población de fragmentos derivados de la biblioteca genómica (s) se hibrida con parcial, o, a menudo ssADN o ARN aproximadamente completa que corresponde a la cadena opuesta. El montaje de genes quiméricos complejos a partir de esta población es entonces mediado por la eliminación a base de nucleasa de extremos del fragmento no híbrido, la polimerización para llenar huecos entre tales fragmentos y la posterior ligadura de cadena sencilla. La hebra de polinucleótido parental puede ser eliminada por digestión (por ejemplo, si ARN o uracilcontaining), separación magnética en condiciones de desnaturalización (si se etiqueta de una manera propicia para tal separación) y otros métodos de separación / purificación disponibles. Alternativamente, la hebra parental opcionalmente se copurifica con las hebras quiméricas y se elimina durante las posteriores etapas de detección y procesamiento. Detalles adicionales con respecto a este enfoque se encuentran, por ejemplo, en "Single-Stranded Nucleic Acid Template Mediated Recombination and Acid Fragment Isolation" por Aftholter, WO 01/64864.

Métodos de recombinación también se pueden realizar en formato digital en un sistema de procesamiento de información. Por ejemplo, los algoritmos pueden ser utilizados en un ordenador para recombinar cadenas de secuencias que corresponden a biomoléculas homólogas (o incluso no homólogas). De acuerdo con realizaciones específicas de la descripción, después del procesamiento en un sistema informático, las cadenas de secuencia resultantes se pueden convertir en ácidos nucleicos mediante síntesis de ácidos nucleicos que corresponden a las secuencias recombinadas, por ejemplo, de acuerdo con técnicas de síntesis / gen de reensamblaje de oligonucleótidos. Este enfoque puede generar variantes aleatorias, parcialmente aleatorias, o diseñadas. Muchos detalles con respecto a diversas realizaciones de recombinación habilitada con ordenador, incluyendo el uso de diversos algoritmos, operadores y similares, en los sistemas informáticos, así como combinaciones de ácidos y / o proteínas (por ejemplo, basándose en la selección del sitio cruzado), así como métodos de recombinación diseñadas, pseudo-aleatorias o aleatorias en el documento WO 00/42560 por Selifonov et al., "Métodos para realizar cadenas de caracteres, los polinucleótidos y polipéptidos que tienen las características deseadas," WO 01/75767 por Gustafsson et a!., "Selección del sitio cruzado in silico," y WO 00/42559 por Selifonov y Stemmer "Métodos de Estructuras de inserción de datos para uso en simulaciones evolutivas."

B. EVOLUCIÓN DIRIGIDA

Evolución dirigida (o alternativamente "evolución artificial") puede llevarse a cabo mediante la práctica de uno o más métodos de generación de diversidad de una manera reiterativa junto con la detección (descrito en más detalle en otra parte en este documento) para generar un conjunto adicional de ácidos nucleicos recombinantes . Por lo tanto, evolución dirigida o artificial puede llevarse a cabo por ciclos repetidos de mutagcncsis y / o recombinación y análisis. Por ejemplo, la mutagénesis y / o la recombinación puede llevarse a cabo en polinucleótidos parentales para generar una biblioteca de polinucleótidos variantes que se expresan a continuación, para generar una biblioteca de variante de proteína que es de pantalla para una actividad deseada. Uno o más proteínas variantes se pueden identificar a partir de la biblioteca variante de la proteína como muestra de mejora en la actividad deseada. Las proteínas identificadas pueden ser traducidas a la inversa para determinar una o más secuencias de polinucleótidos que codifican las variantes de proteínas identificadas, que a su vez se puede mutar o recombinados en una ronda posterior de la generación de diversidad y de cribado.

La evolución dirigida usando formatos basados en la recombinación de la generación de diversidad se describe extensamente en las referencias citadas en este documento. La evolución dirigida usando mutagénesis como la base para la generación de diversidad también es bien conocido en la técnica. Por ejemplo, mutagénesis de conjunto recursiva es un proceso en el que se utiliza un algoritmo para la mutagénesis de proteínas para producir

diversas poblaciones de mutantes fenotípicamente relacionados, miembros de las cuales difieren en la secuencia de aminoácidos. Este método utiliza un mecanismo de retroalimentación para controlar sucesivas rondas de mutagénesis de casete combinatorio. Ejemplos de este enfoque se encuentran en Arkin y Youvan (1992) Proc Natl Acad. Sci USA 89: 7811 - 7815. De manera similar, mutagénesis de conjunto exponencial se puede utilizar para la generación de bibliotecas combinatorias con un alto porcentaje de mutantes únicos y funcionales. Pequeños grupos de residuos en una secuencia de interés se asignaron al azar en paralelo para identificar, en cada posición alterada, aminoácidos que conducen a proteínas funcionales. Ejemplos de tales procedimientos se encuentran en Delegrave y Youvan (1993) Investigación de Biotecnología 11: 1548-1552.

Modelos de estructura-actividad de la presente descripción son útiles para optimizar el proceso de evolución dirigida, independientemente del procedimiento de generación de diversidad empleado. La información derivada de la aplicación de los modelos descritos en este documento puede usarse para diseñar de forma más inteligente bibliotecas realizadas en un proceso de evolución dirigida. Por ejemplo, cuando se desea cambiar o fijar residuos en determinadas posiciones de residuos de aminoácidos, oligonucleótidos sintéticos que incorporan los codones que codifican los residuos de aminoácidos deseados se pueden utilizar en uno de los formatos de recombinación a que se hace referencia en la presente memoria para generar una biblioteca variante de polinucleótido que puede expresarse. Alternativamente, los residuos deseados pueden ser incorporados mediante uno de los diversos métodos de mutagénesis descritos en el presente documento. En cualquier caso, la biblioteca variante de la proteína resultante contendrá variantes de proteínas que incorporan lo que se cree que pueden ser residuos beneficiosos o residuos potencialmente beneficiosos. Este proceso se puede repetir hasta que se identifique una variante de proteína que tiene la actividad deseada.

C. ANÁLISIS / SELECCIÓN PARA ACTIVIDAD

Polinucleótidos de actividad generada en relación con procedimientos de la presente invención se clonan opcionalmente en células para la detección de actividad (o utilizaados en reacciones in vitro de transcripción para realizar productos que posean una pantalla). Además, los ácidos nucleicos se pueden enriquecer, secuenciados, expresados, amplificados in vitro o tratados de cualquier otro método recombinante común.

Los textos generales que describen técnicas de biología molecular útiles en el presente documento, incluyendo la clonación, mutagénesis, construcción de la biblioteca, ensayos de cribado, cultura de célula y similares incluyen Berger y Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology, volumen 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning A Laboratory Manual (2ª Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, Nueva York, 1989 (Sambrook) y Protocolos actuales en Biología Molecular F. M. Ausubel et al., Eds., Current Protocols, una empresa conjunta entre Greene Publishing Associates, Inc. y John Wiley & Sons, Inc., Nueva York (complementado hasta 2000) (Ausubel)). Métodos de transducir células, incluyendo células vegetales y animales, con los ácidos nucleicos son generalmente disponibles, así como métodos para expresar proteínas codificadas por tales ácidos nucleicos. Además de Berger, Ausubel y Sambrook, referencias generales útiles para el cultivo de células animales incluyen Freshney (Culture of Animal Cells, a Manual of Basis Technique, tercera edición Wiley-Liss, Nueva York (1994)) y las referencias citadas en el mismo, Humason (Técnicas de tejido animal cuarta edición W. H. Freeman and Company (1979)) y Ricciardelli, et al., In Vitro Cell Dcv. Biol 25: 1016-24 (1989). Referencias para la clonación de células vegetales, la cultura y la regeneración incluyen Payne et al !. (1992) Plant Cell y Tissue Culture in Liquid Systems, John Wiley & Sons, Inc. Nueva York, NY (Payne); y Gamborg y Phillips (eds) (1995) Plant Cell Tissue Organ Culture: Fundamental Methods, Springer Lab Manual, Springer-Verlag (Berlin Heidelberg Nueva York) (Gamborg). Una variedad de medios de cultivo celular se describen en Atlas y Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL (Atlas). Información adicional para cultivo de células vegetales se encuentra en la literatura comercial disponible, como el Life Science Research Cell Culture Catalogue (1998) de Sigma-Aldrich, Inc (St. Louis, MO) (Sigma-LSRCCC) y, por ejemplo, el Plant Culture Catalogue y el suplemento (1997) también de Sigma-Aldrich, Inc (St. Louis, MO) (Sigma-PCCS).

Los ejemplos de técnicas suficientes para dirigir a los expertos a través de métodos de amplificación in vitro, útil, por ejemplo, para la amplificación de ácidos nucleicos oligonucleótidos recombinados que incluyen reacciones en cadena de la polimerasa (PCR), reacción en cadena de la ligasa (LCR), amplificaciones Qß-replicasa y otras técnicas mediadas por polimerasa ARN (por ejemplo, NASBA). Estas técnicas se encuentran en Berger, Sambrook, y Ausubel, supra, así como en Mullis et al, (1987) Patente de EE.UU. Nº 4.683.202; PCR Protocols A Guide to Methods and Applications (Innis *et al*, eds.) Academic Press Inc. de San Diego, CA (1990) (Innis); Amheim y Levinson (1 de octubre, 1990) C & EN 36-47; The Journal of NIH Research (1991) 3, 81-94; Kwoh et al. (1989) Proc Natl. Acad. Sci USA. 86, 1173; Guatelli et al. (1990) Proc Natl. Acad. Sci EE.UU. 87 de 1874; Lomell et al. (1989) J Clin Chem 35 1826; Landegren et al, (1988) Science 241, 1077-1080.; Van Brunt (1990) Biotechnology 8, 291-294; Wu y Wallace, (1989) Gen 4, 560; Barringer et al. (1990) Gen 89, 117, y Sooknanan y Malek (1995) Biotecnología 13: 563-564. Métodos mejorados de clonación in vitro de ácidos nucleicos amplificados se describen en Wallace et al., patente de EE.UU Nº 5.426.039. Los métodos mejorados de amplificación de ácidos nucleicos grandes por PCR se resumen en Cheng et al. (1994) Nature 369: 684-685 y sus referencias, en los que se generan amplicones de PCR de hasta 40 kb. Un experto apreciará que esencialmente cualquier ARN puede ser convertido en un ADN

bicatenario adecuado para la digestión de restricción, la expansión por PCR y secuenciación usando transcriptasa inversa y una polimerasa. Véase, Ausubel, Sambrook y Berger, *todos supra*.

En un método preferido, las secuencias reensambladas se comprueban para la incorporación de oligonucleótidos de recombinación basados en la familia. Esto se puede hacer mediante la clonación y secuenciación de los ácidos nucleicos, y / o por digestión de restricción, por ejemplo, como se enseña esencialmente en Sambrook, Berger y Ausubel, supra. Además, las secuencias pueden ser amplificadas por PCR y secuenciadas directamente. Por lo tanto, además de, por ejemplo, Sambrook, Berger, Ausubel y Innis (*supra*), metodologías adicionales de secuenciación de PCR son también particularmente útiles. Por ejemplo, la secuenciación directa de amplicones de PCR generado por los nucleótidos resistentes a nucleasa borados selectivamente incorporandos en amplicones durante la PCR y la digestión de los amplicones con una nucleasa para producir fragmentos de tamaño de plantilla se ha realizado (Porter et al (1997) Nucleic Acids Research 25 (8):. 1611-1617). En los métodos, se llevan a cabo cuatro reacciones de PCR en una plantilla, en cada uno de los cuales uno de los trifosfatos de nucleótidos en la mezcla de reacción PCR está parcialmente sustituido con un 2 'deoxinucleósido 5' [P-borano] trifosfato. El nucleótido borado se incorpora estocásticamente en los productos de PCR en diferentes posiciones a lo largo de la amplificación de PCR en un conjunto anidado de fragmentos de PCR de la plantilla. Una exonucleasa que está bloqueada por los nucleótidos incorporados borados se utiliza para escindir los amplicones de PCR. Los amplicones escindidos se separan después por tamaño utilizando electroforesis en gel de poliacrilamida, que proporciona la secuencia del amplicón. Una ventaja de este método es que utiliza un menor número de manipulaciones bioquímicas que con la realización de la secuenciación de estilo Sanger estándar de amplicones de PCR.

Los genes sintéticos son susceptibles de enfoques de clonación y expresión convencionales; por lo tanto, las propiedades de los genes y las proteínas que codifican pueden ser fácilmente examinadas después de su expresión en una célula huésped. Los genes sintéticos también se pueden utilizar para generar productos de polipéptidos de la transcripción y traducción in vitro (libre de células).Los polinucleótidos y polipéptidos de este modo se pueden examinar por su capacidad para unirse a una variedad de ligandos predeterminados, pequeñas moléculas y iones, o polimérico y sustancias heteropoliméricos, incluyendo otras proteínas y epítopos de polipéptidos, así como las paredes celulares microbianas, partículas virales, superficies y membranas .

Por ejemplo, muchos métodos físicos pueden ser utilizados para detectar polinucleótidos que codifican los fenotipos asociados con la catálisis de reacciones químicas ya sea por polynueleotides directamente, o por polipéptidos codificados. Únicamente con el propósito de ilustración, y en función de las características específicas de determinadas reacciones químicas predeterminadas de interés, estos métodos pueden incluir una multitud de técnicas bien conocidas en la técnica, que representan una diferencia física entre el sustrato (s) y producto (s), o los cambios en los medios de reacción asociados con la reacción química (por ejemplo, cambios en las emisiones electromagnéticas, la adsorción, la disipación y de fluorescencia, ya sea UV, visible o infrarroja (calor)). Estos métodos también se pueden seleccionar de cualquier combinación de los siguientes: espectrometría de masas; resonancia magnética nuclear; materiales marcados isotópicamente, partición y métodos espectrales que representan la distribución de isótopos o la formación de producto etiquetado; métodos espectrales y químicas para detectar cambios en ion o composiciones elementales de producto (s) de reacción (incluyendo cambios en el pH, iones inorgánicos y orgánicos y similares) que acompañan. Otros métodos de ensayos físicos adecuados para uso en los métodos de la presente memoria se pueden basar en el uso de biosensores específicos para el producto (s) de reacción, incluyendo los que comprenden anticuerpos con propiedades de reportero, o los basados en el reconocimiento de afinidad in vivo junto con la expresión y la actividad de un gen indicador. Los ensayos enzimáticos de acoplamiento para la detección de producto de reacción y selecciones de muerte celular del crecimiento de la vida in vivo también se pueden usar cuando sea apropiado. Independientemente de la naturaleza específica de los ensayos físicos, todos ellos se utilizan para seleccionar una actividad deseada, o combinación de actividades deseadas, proporcionadas o codificadas por una biomolécula de interés.

El ensayo específico utilizado para la selección dependerá de la aplicación. Se conocen muchos ensayos para proteínas, receptores, ligandos y similares. Los formatos incluyen la unión a componentes inmovilizados, o la viabilidad celular del organismo, producción de composiciones de reportero, y similares.

Ensayos de alto rendimiento son particularmente adecuados para el cribado de bibliotecas empleados en la presente invención. En ensayos de alto rendimiento, es posible cribar hasta varios miles de variantes diferentes en un solo día. Por ejemplo, cada pocillo de una placa de microtitulación puede utilizarse para realizar un ensayo separado, o, si efectos de tiempo de concentración o de incubación se deben observar, cada 5-10 pozos puede probar una sola variante (por ejemplo, a diferentes concentraciones). Por lo tanto, una única placa de microtitulación estándar puede ensayar aproximadamente 100 (por ejemplo, 96) reacciones. Si se utilizan placas de 1536 pocillos, entonces una sola placa puede ensayar fácilmente desde aproximadamente 100 a aproximadamente 1.500 reacciones diferentes. Es posible ensayar varias placas diferentes por día; pantallas de ensayo para hasta aproximadamente 6.000-20.000 ensayos diferentes (es decir, la participación de diferentes ácidos nucleicos, proteínas codificadas, concentraciones, etc.) es posible utilizando los sistemas integrados de la invención. Más recientemente, los enfoques microfluídicos al reactivo de manipulación se han desarrollado, por ejemplo, por Caliper Technologies (Mountain View, CA) que puede proporcionar métodos de ensayo de microfluidos de muy alto rendimiento.

Sistemas de cribado de alto rendimiento están disponibles comercialmente (véase, por ejemplo, Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, etc.) . Estos sistemas típicamente automatizan procedimientos completos, incluyendo todo pipeteado de muestras y reactivos, dispensación de líquido, incubaciones cronometradas, y lecturas finales de la microplaca en detector (s) apropiada para el ensayo. Estos sistemas configurables proporcionan un alto rendimiento y una rápida puesta en marcha, así como un alto grado de flexibilidad y personalización.

Los fabricantes de estos sistemas proporcionan protocolos detallados para diversos ensayos de cribado de alto rendimiento. Así, por ejemplo, Zymark Corp. proporciona boletines técnicos que describen sistemas de exploración para detectar la modulación de la transcripción de genes, la unión del ligando, y similares.

Una variedad de equipos periféricos disponibles en el mercado y el software está disponible para la digitalización, almacenamiento y análisis de un video digitalizado o imágenes ópticas u otro ensayo digitalizadas, por ejemplo, el uso de PC (x86 de Intel chip Pentium o compatible - máquinas basadas en DOS™, OS2™, WINDOWS™, o Windows NT™), ordenadores Macintosh™, o UNIX (por ejemplo, la estación de trabajo Sun ™).

Sistemas de análisis incluyen típicamente un ordenador digital con software para dirigir una o más etapa de uno o más de los métodos del presente documento, y, opcionalmente, también incluyen, por ejemplo, software de control de líquido de alto rendimiento, software de análisis de imágenes, software de interpretación de datos, una armadura de control de líquido robótico para la transferencia de soluciones desde un origen a un destino unido operativamente al ordenador digital, un dispositivo de entrada (por ejemplo, un teclado de ordenador) para introducir datos a la computadora digital para el control de operaciones o líquido de alta rendimiento para transferir la armadura de control de líquido robótico y, opcionalmente, un escáner de imágenes para la digitalización de señales de la etiqueta de componentes de ensayo etiquetados. El escáner de imágenes puede interactuar con el software de análisis de imagen para proporcionar una medición de la intensidad del marcador de la sonda. Típicamente, la sonda de medición de la intensidad etiqueta es interpretada por el software de interpretación de datos para mostrar si la sonda marcada se hibrida con el ADN sobre el soporte sólido.

Recursos de hardware y software computacionales disponibles que pueden emplearse en los métodos de la invención descritos en este documento (por hardware, cualquier sistema Unix del rango de medio precio (por ejemplo, de Sun Microsystems) o al final incluso Macintosh o PC de alta gama serán suficientes).

En algunas formas de realización, las células, las placas virales, esporas o similares, que comprenden en los productos de recombinación mediada por oligonucleótidos in vitro o realizaciones físicas de silico en ácidos nucleicos recombinados, se pueden separar en medios sólidos para producir colonias individuales (o placas). El uso de un selector automático de colonias (por ejemplo, Q-bot, Genetix, Reino Unido), se colonias o placas se identifican, se recogen, y hasta 10.000 diferentes mutantes se inocularon en placas de microtitulación de 96 pocillos que contienen dos bolas de vidrio de 3 mm / pocillo. La Q-bot no recoge toda una colonia sino más bien inserta un pasador a través del centro de la colonia y sale con una pequeña muestra de células, (o micelio) y esporas (o virus en aplicaciones de placa). El tiempo que el pasador está en la colonia, el número de bajadas para inocular el medio de cultivo, y el tiempo que la clavija esté en ese medio afectan el tamaño del efecto inóculo, y cada parámetro puede ser controlado y optimizado.

El proceso uniforme de recogida de colonia automatizada como la Q-bot disminuye el error de manipulación humana y aumenta la velocidad de establecimiento de cultivos (aproximadamente 10.000 / 4 horas). Estos cultivos se agitan opcionalmente en una temperatura y humedad incubadora controlada. Bolas de cristal opcionales en las placas de microtitulación actúan para promover la aireación uniforme de las células y la dispersión de fragmentos celulares (por ejemplo, de micelio) similares a las cuchillas de un fermentador. Los clones procedentes de cultivos de interés pueden ser aislados por dilución limitante. Como también se describe supra, placas o células que constituyen bibliotecas también se pueden cribar directamente para la producción de proteínas, ya sea mediante la detección de la hibridación, actividad de la proteína, la proteína de unión a anticuerpos, o similares. Para aumentar las posibilidades de identificar una piscina de tamaño suficiente, una preselección que aumenta el número de mutantes procesados por 10 veces se puede utilizar. El objetivo de la pantalla principal es identificar rápidamente los mutantes que tienen títulos de productos iguales o mejores que la cepa padre (s) para mover sólo estos mutantes a cultivo celular líquido para su posterior análisis.

Un enfoque para la detección de diversas bibliotecas es la utilización de un procedimiento en fase sólida paralela masiva para detectar células que expresan variantes de polinucleótidos, por ejemplo, polinucleótidos que codifican variantes de la enzima. Aparatos de detección en fase sólida masivamente paralela mediante la absorción, fluorescencia, o FRET están disponibles. Véase, por ejemplo, la patente de EE.UU. No. 5.914.245 a Bylina, et al. (1999); véase también, http: //www.kairos-scientific.com/ Youvan et al. (1999) "Microespectrofotómetro de imagen de fluorescencia (FIMS)" Biotecnología et alia <www.et-al.com> 1: 1-16; Yang et al. (1998) "Imágenes de Alta Resolución Microscopio (HIRIM)" Biotecnología et alia <www.et-al.com> 4: 1-20; y Youvan et al. (1999) "Calibración de transferencia de energía de resonancia de fluorescencia en el uso de derivados de Microscopía GEP genéticamente modificados en los granos de níquel quelantes", publicados en www.kairos-scientific.com. Tras el cribado mediante estas técnicas, las moléculas de interés se aíslan típicamente y opcionalmente secuencian

utilizando métodos que están bien conocidos en la técnica. La información de la secuencia se utiliza entonces como se establece en el presente documento para diseñar una nueva biblioteca de variante de la proteína.

Del mismo modo, un número de sistemas robóticos bien conocidos también se han desarrollado para la química en fase de solución útiles en sistemas de ensayo. Estos sistemas incluyen estaciones de trabajo automáticas como el aparato de síntesis automatizada desarrollado por Takeda Chemical Industries, LTD. (Osaka, Japón) y muchos sistemas robóticos que utilizan brazos robóticos (Zymate II, Zymark Corporation, Hopkinton, Mass.; Orca, Beckman Coulter, Inc. (Fullerton, CA)) que imitan las operaciones sintéticas manuales realizadas por un científico. Cualquiera de los dispositivos anteriores son adecuados para su uso con la presente invención, por ejemplo, para el cribado de alto rendimiento de moléculas codificadas por los ácidos nucleicos desarrollados como se describe en el presente documento. La naturaleza y ejecución de modificaciones a estos dispositivos (si los hay) de modo que puedan funcionar como se analiza en este documento será evidente para los expertos en la técnica relevante.

VII. APARATOS Y SISTEMAS DIGITALES

Como debe ser evidente, las realizaciones de la presente invención emplean procesos que actúan bajo el control de instrucciones y / o datos almacenados en o transferidos a través de uno o más sistemas informáticos. Las realizaciones de la presente invención se refieren también a un aparato para la realización de estas operaciones. Tal aparato puede ser especialmente diseñado y / o construido para los fines requeridos, o puede ser un ordenador de propósito general activado o reconfigurado por un programa de ordenador y / o estructura de datos almacenados en el ordenador de forma selectiva. Los procesos que se presentan en este documento no están intrínsecamente relacionados con ningún equipo en particular u otro aparato. En particular, varias máquinas de propósito general pueden usarse con programas escritos de acuerdo con los métodos de este documento. En algunos casos, sin embargo, puede ser más conveniente construir un aparato especializado para realizar las operaciones del método requeridas. Una estructura particular para una variedad de estas máquinas aparecerán a partir de la descripción dada a continuación.

Además, las realizaciones de la presente invención se refieren a los productos de los medios de comunicación o de programa de ordenador legibles por ordenador que incluyen instrucciones de programas y / o datos (incluidas estructuras de datos) para la realización de diversas operaciones implementadas en ordenador. Ejemplos de medios legibles por ordenador incluyen, pero no se limitan a medios magnéticos tales como discos duros, disquetes, cintas magnéticas; medios ópticos, tales como dispositivos de CD-ROM y dispositivos holográficos; medios magneto-ópticos; dispositivos de memoria de semiconductores y dispositivos de hardware que están especialmente configurados para almacenar y ejecutar instrucciones de programa, como dispositivos de memoria de sólo lectura (ROM) y memoria de acceso aleatorio (RAM), y, a veces circuitos integrados de aplicación específica (ASIC), dispositivos lógicos programables (PLDs) y medios de transmisión de señales para la entrega de instrucciones legibles por ordenador, tales como redes de área local, redes de área amplia, e Internet. Las instrucciones de datos y de programas de esta invención también pueden ser incluidas en una onda portadora u otro medio de transporte (por ejemplo, líneas ópticas, líneas eléctricas, y / o ondas de radio).

Ejemplos de instrucciones de programa incluyen tanto el código de bajo nivel, como el producido por un compilador y archivos que contienen código de nivel superior que pueden ser ejecutados por el ordenador usando un intérprete. Además, las instrucciones de programa incluyen un código máquina, código fuente y cualquier otro código que controla directamente o indirectamente la operación de una máquina de computación de acuerdo con esta invención. El código puede especificar la entrada, salida, cálculos, condicionales, ramas, bucles iterativos, etc.

Aplicaciones de escritorio estándar, tales como procesadores de texto (por ejemplo, Microsoft WordTM o Corel WordPerfectTM) y el software de base de datos (por ejemplo, el software de hoja de cálculo como Microsoft Excel TM, Corel Quattro Pr0TM, o programas de bases de datos como Microsoft AccessTM o ParadoxTM) se pueden adaptar a la presente invención mediante la introducción de una o más cadenas de caracteres en el software que se carga en la memoria de un sistema digital, y la realización de una operación como se ha señalado en el presente documento en la cadena de caracteres. Por ejemplo, los sistemas pueden incluir el software anterior que tiene la información de la cadena de caracteres adecuada, por ejemplo, utilizándose en conjunción con una interfaz de usuario (por ejemplo, una interfaz gráfica de usuario en un sistema operativo estándar tal como un Windows, Macintosh o un sistema Linux) para manipular cadenas de caracteres. Programas de alineación especializados como CACHARRO y BLAST también se pueden incorporar en los sistemas de la invención, por ejemplo, para la alineación de los ácidos nucleicos o proteínas (o cadenas de caracteres correspondientes) como un paso preparatorio para la realización de una operación en cualesquiera secuencias alineadas. El software para realizar PCA (por ejemplo, como está disponible comercialmente de Partek) u otras operaciones estadísticas también pueden ser incluidas en el sistema digital.

Los sistemas típicamente incluyen, por ejemplo, un ordenador digital con el software para la alineación y la manipulación de secuencias de acuerdo con las operaciones observadas en el presente documento, o para la realización de PCA, análisis de redes neuronales o similares, así como los conjuntos de datos introducidos en el sistema de software que comprende secuencias u otros datos que se asigna o se manipula. El ordenador puede ser,

por ejemplo, un PC (x86 de Intel o DOS™ chip Pentium compatible, OS2™, WINDOWS™, Windows NT™, WINDOWS95™, WINDOWS98™, compatible con Apple, compatible con Macintosh™, LINUX, compatible con Power PC o un compatible con UNIX (por ejemplo, estación de trabajo o máquina SUN™) u otro equipo disponible en el mercado común que es conocido para un experto. Software para la alineación de secuencias o para su manipulación puede ser construido por un experto usando un lenguaje de programación estándar, tal como VisualBasic, Fortran, Basic, Java, o similares, de acuerdo con los métodos de la presente memoria.

Cualquier controlador u ordenador opcionalmente incluye un monitor que puede incluir, por ejemplo, un tubo de rayos catódicos ("CRT") pantalla, una pantalla plana (por ejemplo, pantalla de matriz activa de cristal líquido, pantalla de cristal líquido), u otros. Circuitos de ordenador a menudo se colocan en una caja que incluye numerosos chips de circuito integrado, tal como un microprocesador, memoria, circuitos de interfaz, y otros. La caja también opcionalmente incluye una unidad de disco duro, una unidad de disquete, una unidad extraíble de alta capacidad, como un CD-ROM grabable, y otros elementos periféricos comunes. Dispositivos de entrada tales como un teclado o un ratón opcionalmente permiten la entrada de un usuario y la selección de secuencias que se comparan o manipulan de otra manera en el sistema informático correspondiente.

El ordenador típicamente incluye software apropiado para la recepción de instrucciones del usuario, ya sea en la forma de la entrada del usuario en un conjunto de campos de parámetro, por ejemplo, en un GUI, o en la forma de instrucciones preprogramadas, por ejemplo, preprogramadas para una variedad de diferentes operaciones específicas. Entonces, el programa convierte estas instrucciones a un lenguaje apropiado para instruir al sistema para llevar a cabo cualquier operación deseada. Por ejemplo, además de realizar manipulaciones estadísticas de espacio de datos, un sistema digital puede dar instrucciones a un sintetizador de oligonucleótidos para sintetizar oligonucleótidos para la reconstrucción de genes, o incluso a la orden de fuentes comerciales de oligonucleótidos (por ejemplo, mediante la impresión de los formularios de pedido apropiadas o mediante el enlace a un formulario de pedido en el Internet).

El sistema digital también puede incluir elementos de salida para el control de la síntesis de ácido nucleico (por ejemplo, basado en una secuencia o una alineación de una secuencia de la presente memoria), es decir, un sistema integrado de la invención incluye opcionalmente un sintetizador de oligonucleótidos o un controlador de la síntesis de oligonucleótidos. El sistema puede incluir otras operaciones que tienen lugar aguas abajo de una alineación o de otra operación realizada mediante una cadena de caracteres correspondiente a una secuencia en el presente documento, por ejemplo, como se señaló anteriormente con referencia a ensayos.

En un ejemplo, el código que contiene los métodos de la invención se realiza en un medio fijo o componente de programa transmisible que contienen instrucciones de lógica y / o datos que cuando se carga en un dispositivo informático configurado apropiadamente hace que el dispositivo lleve a cabo un operador genético en uno o más cadenas de caracteres. La figura 16 muestra un dispositivo digital de ejemplo 2200 que debe ser entendido como un aparato lógico que pueda leer las instrucciones de los medios de comunicación 2217, puerto de red 2219, usuario del teclado de entrada 2209, la entrada del usuario 2211 u otros medios de entrada. El aparato 2200 a partir de entonces puede usar esas instrucciones para dirigir las operaciones estadísticas en el espacio de datos, por ejemplo, para la construcción de 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 encarnar la invención es un sistema de ordenador como en el sistema de ordenador 2200, que comprende CPU 2207, teclado de dispositivos opcional de entrada de usuario 2209, y un dispositivo señalador GUI 2211, así como componentes periféricos tales como unidades de disco 2215 y monitorear 2205 (que muestra cadenas de caracteres modificadas GO y proporciona la selección simplificada de subconjuntos de dichas cadenas de caracteres por un usuario. Medios fijos 2217 se utilizan opcionalmente para programar el sistema global y pueden incluir, por ejemplo, un tipo de disco de medios ópticos o magnéticos u otro elemento de almacenamiento de memoria electrónica. Un puerto de comunicación 2219 se puede utilizar para programar el sistema y puede representar cualquier tipo de conexión de comunicación.

La invención también se puede realizar dentro de los circuitos de un circuito integrado para aplicaciones específicas (ASIC) o dispositivo lógico programable (PLD). En tal caso, la invención se realiza en un lenguaje descriptor legible por ordenador que se puede utilizar para crear un ASIC o PLD. La invención también se puede realizar dentro de los procesadores de circuitos lógicos o de una variedad de otros aparatos digitales, tales como PDAs, ordenadores, pantallas de ordenador portátil, equipo de edición de imágenes, etc.

En un aspecto preferido, el sistema digital comprende un componente de aprendizaje donde se controlan los resultados de programas de físicos de montaje oligonucleótido (composiciones, abundancia de los productos, procesos diferentes), en relación con los ensayos físicos, y se establecen correlaciones. Combinaciones exitosas y fallidas están documentadas en una base de datos para proporcionar justificación / preferencias de usuario-base o sistema digital de selección basado en conjuntos de parámetros para los procesos posteriores descritos en el presente documento, los cuales implican el mismo conjunto de cadenas de caracteres parentales / ácidos nucleicos / proteínas (o incluso secuencias no relacionadas, donde la información proporciona información de mejora de procesos). Las correlaciones se utilizan para modificar los procesos posteriores de la invención, por ejemplo, para optimizar el proceso particular. Este ciclo de síntesis física, selección y correlación se repite opcionalmente para

optimizar el sistema. Por ejemplo, una red neuronal de aprendizaje se puede utilizar para optimizar los resultados.

VIII. REALIZACIONES EN LAS PÁGINAS WEB

El Internet incluye ordenadores, aparatos de información, y redes ordenador que se interconectan a través de enlaces de comunicación. Los ordenadores interconectados intercambian información a través de diversos servicios, como el correo electrónico, FTP, la World Wide Web ("WWW") u otros servicios, incluyendo servicios seguros. Se puede entender que el servicio WWW permite un sistema de servidor de ordenador (por ejemplo, un servidor Web o un sitio web) para enviar páginas web de información a un aparato de información de cliente remoto o sistema informático. El sistema informático del cliente remoto puede mostrar las páginas web. En general, cada recurso (por ejemplo, ordenador o página web) de la WWW es identificable únicamente por un Localizador Uniforme de Recursos ("URL"). Para ver o interactuar con una página web específica, un sistema informático cliente especifica una dirección de URL de esa página web en una solicitud. La solicitud se reenvía a un servidor que soporta esa página web. Cuando el servidor recibe la solicitud, envía la página web para el sistema de información del cliente. Cuando el sistema informático cliente recibe la página web, se puede mostrar la página web usando un navegador o puede interactuar con la página web o interfaz que se disponga otra cosa. Un navegador es un módulo lógico que afecta la solicitud de páginas web y la exhibición o interración con las páginas web.

En la actualidad, las páginas web que se pueden visualizar se definen típicamente usando un lenguaje de marcado de hipertexto ("HTML"). HTML proporciona un conjunto estándar de etiquetas que definen cómo una página web se va a mostrar. Un documento HTML contiene varias etiquetas que controlan la visualización de textos, gráficos, controles, y otras características. El documento HTML puede contener URLs de otras páginas web disponibles en dicho sistema informático de servidor u otros sistemas informáticos de servidor. URLs también pueden indicar otros tipos de interfaces, incluyendo cosas tales como scripts CGI o interfaces ejecutables, que utilizan los dispositivos de información para comunicarse a distancia con dispositivos de información o servidores sin mostrar necesariamente información a un usuario.

El Internet es especialmente propicio para proporcionar servicios de información a uno o más clientes remotos. Los servicios pueden incluir elementos (por ejemplo, la música o las cotizaciones de bolsa) que se entregan por vía electrónica a un comprador a través de Internet. Los servicios pueden incluir también órdenes de manipulación de objetos (por ejemplo, tiendas de comestibles, libros, o compuestos químicos o biológicos, etc.) que pueden ser entregados a través de canales convencionales de distribución (por ejemplo, una empresa de transporte público). Los servicios también pueden incluir el manejo de pedidos de artículos, tales como reservas de avión o de teatro, que tienen acceso a un comprador en un momento posterior. Un sistema informático servidor puede proporcionar una versión electrónica de una interfaz que muestra los elementos o servicios que están disponibles. Un usuario o un comprador potencial puede acceder a la interfaz usando un navegador y seleccionar varios elementos de interés. Cuando el usuario ha terminado de seleccionar los elementos deseados, el sistema informático servidor puede solicitar al usuario la información necesaria para completar el servicio. Esta información de la orden específica de la transacción puede incluir el nombre del comprador o cualquier otra identificación, una identificación para el pago (por ejemplo, un número de orden de compra o número de cuenta corporativa), o información adicional necesaria para completar el servicio, tales como información de vuelo. Bases de datos y software NCBI.

Entre los servicios de particular interés que se pueden proporcionar en Internet y a través de otras redes son datos biológicos y bases de datos biológicos. Tales servicios incluyen una variedad de servicios prestados por el Centro Nacional de Información Biotecnológica (NCBI) de los Institutos Nacionales de Salud (NIH). NCBI se encarga de la creación de sistemas automatizados para el almacenamiento y análisis de los conocimientos sobre la biología molecular, la bioquímica y la genética; la facilitación del uso de este tipo de bases de datos y software de la investigación y la comunidad médica; la coordinación de los esfuerzos para reunir información biotecnologica tanto a nivel nacional como internacional; y la realización de estudios sobre los métodos avanzados de procesamiento de la información basado en ordenador para el análisis de la estructura y función de las moléculas biológicamente importantes.

NCBT tiene la responsabilidad de la base de datos GenBank de secuencia de ADN. La base de datos se ha construido a partir de secuencias presentadas por los laboratorios individuales, y el intercambio de datos con las bases de datos internacionales de secuencias de nucleótidos, el Laboratorio Europeo de Biología Molecular (EMBL) y la Base de datos de ADN de Japón (DDBJ), e incluye datos de la secuencia de patentes presentadas ante la Oficina de Patentes y Marcas de EE.UU. Además de GenBank, NCBI soporta y distribuye una gran variedad de bases de datos para las comunidades médicas y científicas. Estos incluyen la herencia mendeliana en el hombre (OMIM), la base de datos de Modelado Molecular (MMDB) de las estructuras 3D de proteínas, la Colección de Secuencia Única de Genes Humanos (UniGene), un mapa de genes del genoma humano, el Navegador sobre Taxonomía y el Cancer Genome Anatomy Project (CGAP), en colaboración con el Instituto Nacional del Cáncer. Entrez es el sistema de búsqueda y recuperación de NCBI que proporciona a los usuarios acceso integrado a la secuencia, la cartografía, la taxonomía, y los datos estructurales. Entrez también proporciona vistas gráficas de secuencias y mapas de cromosomas. Una característica de Entrez es la capacidad de recuperar secuencias relacionadas, estructuras y referencias. BLAST, como se describe en el presente documento, es un programa para la búsqueda de similitud de secuencia desarrollado en NCBI para la identificación de genes y las características

genéticas que pueden ejecutar búsquedas de secuencia contra la base de datos de ADN. Herramientas de software adicionales proporcionadas por NCBI incluyen: Open Reading Frame Finder (ORF Finder), PCR electrónica, y las herramientas de envío de secuencias, de lentejuelas y BankIt. Las diversas bases de datos de NCBI y herramientas de software están disponibles en la WWW o por FTP o por los servidores de correo electrónico. Más información está disponible en www.ncbi.nlm.nih.gov.

Algunos datos biológicos disponibles en internet son datos que se ve generalmente con un navegador especial "plug-in" u otro código ejecutable. Un ejemplo de un sistema de este tipo es CHIME, un plug-in de navegador que permite una visualización de tres dimensiones virtual interactivo de estructuras moleculares, incluyendo las estructuras moleculares biológicas. Más información sobre la función está disponible en www.mdlchime.com/chime/. Oligos en línea, Gene, u Ordenación de Proteínas

Una variedad de compañías e instituciones proporcionan los sistemas en línea para ordenar compuestos biológicos. Ejemplos de tales sistemas se pueden encontrar en

www.genosys.com/oligo_custinfo.cfm o www.genomictechnologies.com/Qbrowser2_FP.html. Por lo general, estos sistemas aceptan algún descriptor de un compuesto biológico deseado (tal como un oligonucleótido, cadena de ADN, cadena de ARN, secuencia de aminoácidos, etc.) y después el compuesto deseado se fabrica y se envía al cliente en una solución líquida u otra forma apropiada. Para ilustrar adicionalmente, los métodos de esta invención pueden implementarse en un entorno informático localizado o distribuido. En un entorno distribuido, los métodos pueden ser implementados en un único ordenador que comprende varios procesadores o en una multiplicidad de ordenadores. Los ordenadores pueden estar vinculados, por ejemplo, a través de un bus común, pero más preferiblemente el equipo (s) son nodos de una red. La red puede ser una red de área amplia local o generalizada o dedicada y, en ciertas realizaciones preferidas, los ordenadores pueden ser componentes de una intranet o un internet.

En una realización de Internet, un sistema de cliente normalmente ejecuta un navegador Web y está acoplado a un equipo servidor de ejecución de un servidor Web. El navegador web es típicamente un programa como el Explorador de Internet de IBM, Internet Explorer de Microsoft, Netscape, Opera o Mosaic. El servidor Web es típicamente, pero no necesariamente, un programa como el daemon de HTTP de IBM u otro daemon www (por ejemplo, formularios basados en Linux del programa). El equipo cliente se acopla bi-direccionalmente a un ordenador servidor a través de una línea o a través de un sistema inalámbrico. A su vez, el equipo servidor es bidireccional, junto con una página web (servidor que aloja el sitio web) que proporciona acceso al software de la aplicación de los métodos de esta invención.

Como se ha mencionado, un usuario de un cliente conectado a Intranet o Internet puede provocar que el cliente solicite recursos que son parte del sitio (s) web que aloja la aplicación (s) que proporciona una implementación de los métodos de esta invención. El programa (s) de servidor luego procesa el pedido de devolución de los recursos especificados (suponiendo que están actualmente disponibles). La convención de nomenclatura estándar (es decir, Localizador Uniforme de Recursos ("URL")) abarca varios tipos de nombres de ubicaciones, en la actualidad, incluyendo subclases como el Protocolo de Transferencia de Hipertexto ("http"), File Transport Protocol ("ftp"), Gopher, y Wide Area Information Services ("WAIS"). Cuando se descargue un recurso, puede incluir las direcciones URL de recursos adicionales. Por lo tanto, el usuario del cliente puede aprender fácilmente de la existencia de nuevos recursos que él o ella no había pedido específicamente.

El software que implementa el método (s) de esta invención se puede ejecutar de forma local en el servidor que aloja el sitio web en una arquitectura cliente-servidor cierto. Por lo tanto, el ordenador del cliente envía solicitudes al servidor de huésped que ejecuta el proceso solicitado (s) a nivel local y luego descarga los resultados de vuelta al cliente. Alternativamente, los métodos de esta invención pueden implementarse en un formato "multi-nivel" en la que un componente del método (s) se llevan a cabo de forma local por el cliente. Esto se puede implementar por software descargado desde el servidor a petición del cliente (por ejemplo, una aplicación Java) o puede ser implementada por software "permanentemente" instalado en el cliente.

En una realización, la aplicación (s) que implementa los métodos de esta invención se divide en tramas. En este paradigma, es útil ver una aplicación no tanto como un conjunto de características o funciones pero, en cambio, como una colección de marcos discretos o puntos de vista. Una aplicación típica, por ejemplo, generalmente incluye un conjunto de elementos de menú, cada uno de los cuales invoca un marco en particular - es decir, una forma que manifiesta cierta funcionalidad de la aplicación. Con esta perspectiva, una aplicación es vista no como un cuerpo monolítico de código, sino como un conjunto de applets, o paquetes de funcionalidad. De esta manera desde un navegador, un usuario seleccionaría un enlace de página Web que, a su vez, invoca un marco en particular de la aplicación (es decir, una sub-aplicación). Así, por ejemplo, una o más tramas pueden proporcionar la funcionalidad para la introducción y / o que codificación molecular (s) biológica en uno o más espacios de datos, mientras que otro marco proporciona herramientas para refinar un modelo del espacio de datos.

En ciertas realizaciones, los métodos de esta invención se implementan como una o más tramas que proporciona, por ejemplo, la siguiente functionalidad (es). Función (s) para codificar dos o más moléculas biológicas en cadenas

de caracteres para proporcionar una colección de dos o más diferentes cadenas de caracteres iniciales, en el que cada una de dichas moléculas biológicas comprende un conjunto seleccionado de subunidades; funciones para seleccionar al menos dos subseries de las cadenas de caracteres; funciones para concatenar las subseries para formar uno o más productos cuerdas alrededor de la misma longitud que una o más de las cadenas de caracteres iniciales; funciones para añadir (colocar) las cadenas de productos de una colección de cadenas y funciones para poner en práctica alguna de las funciones establecidas en el presente documento.

Las funciones para la distribución de dos o más moléculas biológicas en el espacio de datos puede proporcionar una o más ventanas en el que el usuario puede insertar representación (s) de moléculas biológicas. Además, la función de codificación también, opcionalmente, proporciona acceso a bases de datos privadas y / o públicas accesibles a través de una red local y / o la intranet con lo cual una o más secuencias contenidas en las bases de datos se pueden introducir en los métodos de esta invención. Así, por ejemplo, en una realización, donde el usuario extremo introduce un ácido nucleico secuenciado en la función de codificación, el usuario puede, opcionalmente, tener la capacidad de solicitar una búsqueda de GenBank® y la entrada de una o más de las secuencias devueltas por una búsqueda de este tipo en la codificación y / o función de diversidad de la generación.

Los métodos de aplicación de Intranet y / o Intranet realizaciones de procesos de acceso computacionales y / o de datos son bien conocidos por los expertos en la materia y están documentados en gran detalle (véase, por ejemplo, Cluer et al. (1992) "un marco general para la optimización de las consultas orientadas a objetos," Proc SIGMOD Conferencia Internacional sobre la Gestión de Datos, San Diego, California, junio. 2-5, 1992, SIGMOD Record, vol. 21, No. 2, Jun., 1992; Stonebraker, M., editor; ACM Press, pp 383-392.; ISO-ANSI, Borrador de Trabajo, "Tecnología de la Información Base de Datos de Lenguaje SQL, "Jim Melton, Editor, Organización Internacional para la Estandarización y el Instituto Nacional Americano de Estándares, Jul., 1992; Microsoft Corporation, "ODBC 2.0 referencia del programador y guía SDK. La base de datos de Microsoft de estándar abierto para Microsoft Windows™ y Windows NT™, Microsoft Open Database Connectivity™. Equipo de desarrollo de software", 1992, 1993, 1994 Microsoft Press, pp 3-30 y 41-56; Borrador de Trabajo ISO,." Base de datos Lenguaje SQL-Parte 2: Fundación (SQL / Fundación), "CD9075-2: 199.chi .SQL, Sep. 11, 1997 y similares). pormenores relevantes con respecto a las aplicaciones basadas en la web se encuentran en el documento WO 00/42559, titulado "MÉTODOS DE POBLAR ESTRUCTURAS DE DATOS PARA SU USO EN SIMULACIONES EVOLUTIVAS", por Selifonov y Stemmer.

IX. EJEMPLOS - IDENTIFICACIÓN FUNCIONAL DE LIMITACIONES EN PROTEÍNAS POR LA BARAJADA DE ADN SINTÉTICO

El siguiente ejemplo no limitante se ofrece sólo a modo de ilustración.

La evolución de proteínas se manifiesta por cambios de aminoácidos en la codificación de secuencia. Estos cambios de aminoácidos están limitados por la presión selectiva continua para la función, lo que desemboca en cambios independientes y correlacionados en descendientes de proteína. En esta sección se presenta un método para diferenciar covariación entre aminoácidos, que reflejan selección funcional, de la covariación que simplemente resulta de un origen ancestral común.

El cribado funcional y secuenciación de secuencias sugiere que la mayoría de la covariación observada en secuencias de origen natural resulta de descenso filogenético en lugar de limitaciones funcionales. Las covariaciones funcionales que se identifican se encuentran principalmente en los elementos estructurales locales, pero también hay algo de covariación que ocurre en distancias más largas en genes / proteínas. En general, los genes y las proteínas son muy plásticos y han evolucionado para minimizar la interdependencia de los cambios de aminoácidos permitidos para facilitar la adaptación.

Durante la evolución divergente, las secuencias de proteínas cambian mientras que la función bioquímica de la proteína generalmente se mantiene. Cambio correlacionado entre residuos funcionalmente vinculados en una proteína prevé la conservación de la estructura y función de proteínas a lo largo del proceso evolutivo. El vínculo funcional entre los residuos de covariación puede deberse, por ejemplo, al contacto estructural o un efecto indirecto a través de interacciones con sustratos, productos, cofactores u otras proteínas. Mutaciones independientes entre residuos funcionalmente vinculados a menudo son desventajosos, pero dos mutaciones simultáneas pueden permitir que la proteína de función se conserva. Alternativamente, dos o más residuos pueden covariar simplemente debido a un origen ancestral común. Herramientas analíticas actuales están limitadas en la capacidad de separar lo funcional de la covariación filogenética (ancestral) en una familia de proteínas ortólogas. Herramientas estadísticas se limitan tanto por la cantidad de datos para inferir covariación y también se encuentra limitado por los modelos evolutivos para la explicación de los datos. Véase, Wollenberg, K. R. & Atchley, W. R. La separación de las asociaciones filogenéticas y funcionales en secuencias biológicas mediante el uso de la rutina de carga paramétrica. Proc Nat'lAcad. Sci 97, 3288-91. (2000); Gaucher, E. A., Miyamoto, M. y M. Benner, S. A. análisis de estructura-función de las proteínas, utilizando enfoques evolutivos basados en covarion: Factores de elongación. Proc Acad. Sci 98, 548-552 (2001); Larson, S. M., Di Nardo, A. A. & Davidson, A. R. El análisis de la covariación en una alineación de secuencias de dominio SH3: aplicaciones en la predicción de contacto terciario y el diseño de

compensación de sustituciones básicas hidrófobas. J Mol Biol 303, 433-46. (2000); Pollock, D. D., Taylor, W. R. & Goldman, N. Residuos de proteínas coevolucionados: la identificación máxima de verosimilitud y su relación con la estructura. J Mol Biol 287, 187-98. (1999; y Atchley, WR, Wollenberg, KR, Fitch, WM, Terhalle, W. y Vestimenta, AW Las correlaciones entre los sitios de aminoácidos en los dominios de proteínas bHLH: un análisis teórico de información Mol Biol Evol 17, 164-78 (2000).

Si mutaciones puntuales secuenciales son el principal mecanismo para la evolución divergente, la mayoría de los cambios de aminoácidos deben ocurrir de forma independiente: dos mutaciones simultáneas serán extremadamente poco comunes (por ejemplo, a razón de una mutación por cada 10^9 pares de bases para una sola división celular en E. coli).

Aquí se describe un experimento en el que todos los aminoácidos en una familia de proteínas se desacoplan deliberadamente por ADN sintético arrastrando los pies (es decir, la recombinación de oligonucleótidos sintéticos que corresponden colectivamente en secuencia a un conjunto de polinucleótidos parentales). Al permitir que todos los residuos se varíen independientemente del contexto y luego la detección de función, cualquier covariación derivada de un origen ancestral común se elimina y sólo la covariación que contribuye a la función se mantiene. Las variantes funcionales son analizadas utilizando la teoría de la información mutua para evaluar la covariación entre los residuos. La mayor parte de la covariación observada entre las secuencias parentales no se conserva en las proteínas quiméricas funcionales, indicando que es principalmente una medida de ascendencia ancestral común. Los métodos también identifican residuos de covariación que no se ven entre los parentales debido a los efectos de muestreo.

La barajada sintética se puede realizar en un método independiente de homología que permite que una probabilidad esencialmente igual de cada residuo permitido en cualquier posición dada pueda incorporarse en el producto final. Véase, por ejemplo, el documento WO 00/42561 por Crameri et al., "Recombinación por ácido nucleico de oligonucleótidos mediados" y Ness, J., 20 Minshull, J. & Kim, S. Synthetic Shuffling. Nature Biotech Submitted (2001)). Esto está en contraste con muchos otros formatos de recombinación, donde la distribución de cualquier residuo solo depende de su abundancia y el contexto entre los genes parentales. La barajada sintética resulta en una biblioteca de secuencias que son completamente quiméricas en el nivel de residuos individuales y rica en diversidad natural.

A pesar del gran tamaño total de las bibliotecas que pueden ser generados por redistribución sintética, la caracterización de sólo un pequeño subconjunto de la biblioteca es suficiente para poner a prueba un número significativo de pares de residuos covariados para la correlación con la función. Cualquier par de residuos de aminoácidos covariados se muestrea varias veces entre las variantes completamente caracterizadas. Bibliotecas generadas a través de barajado sintético son una excelente fuente imparcial de datos para analizar la importancia relativa de la covarianza y su distribución en un sistema biológico.

La caracterización de la distribución de una biblioteca pre-seleccionada permite la normalización de la covariación encontrada entre las variantes activas a la distribución inherente de la covarianza de la biblioteca. Cualquier información mutua espuria derivada de una biblioteca imperfecta (por ejemplo sesgos de degeneración de oligonucleótidos producidos durante la síntesis) puede ser eliminada. En general, no hay ninguna, o muy poco, diferencia en la distribución de la diversidad de secuencias entre las variantes pre-seleccionadas y activas. En ambos casos, las variantes se distribuyen de manera uniforme, lo que sugiere ningún sesgo significativo hacia la diversidad procedente de cualquier parental o un grupo de parentales dados. Esto demuestra que las nuevas regiones del espacio de secuencias pueden ser exploradas para la actividad funcional mediante la distribución de las variantes caracterizadas de forma homogénea en la misma secuencia de espacio cubierto por genes parentales. La distancia de secuencia recorrida usando técnicas de evolución dirigida clásicas tales como la mutagénesis aleatoria se limita generalmente a 1-3 residuos de aminoácidos por gen por redonda. La mayoría de las soluciones encontradas a través de barajado sintéticos son en consecuencia inaccesibles por mutagénesis aleatoria.

La covariación entre los residuos inferidos a partir de datos de secuencias biológicas se puede atribuir a cualquiera de las limitaciones funcionales o relaciones filogenéticas. Dado que generalmente no se conoce el origen histórico de las secuencias en cuestión (por lo menos donde las secuencias son de origen natural), no se puede desenrevesar el carácter de covariante de los residuos que participan. Este problema normalmente se ha abordado ya sea a través de la recogida de tantas secuencias como sea posible en virtud de un determinado nodo en un árbol filogenético, o mediante simulaciones por ordenador de posibles caminos evolutivos utilizando un modelo de secuencia de evolución. Ambos enfoques tienen complicaciones e inconvenientes significativos. Una complicación inherente del primer tipo de análisis de covariación es la inclusión de secuencias que han divergido no sólo en mutaciones neutras, sino también en función. La divergencia puede ser pequeño, como en evolución para un óptimo de pH ligeramente diferente, o grande como en evolución para catalizar una reacción relacionada pero diferente. Ninguna enzima ortóloga se ha desarrollado para exactamente las mismas condiciones fisiológicas. Incluso secuencias en el análisis de la covariación que han divergido en función añade ruido a las correlaciones, ya que están sometidas a diferentes presiones selectivas. Otra, tal vez más grave preocupación, es la incapacidad para reunir todas las secuencias bajo un nodo filogenético para asegurar que la distribución en el conjunto de datos es imparcial debido a los efectos de muestreo. En una biblioteca producida por barajado sintético, toda covariación

inherente se retira y la diversidad de aminoácido que se produce en cualquier posición tiene la misma probabilidad de ocurrir en cualquier variante. El cribado de una biblioteca de este tipo (por ejemplo, in vitro) para una función bioquímica definida, identifica toda la covariación derivada de las limitaciones funcionales requeridas para la actividad biológica ensayada de la enzima. El resto de la covariación encontrada entre los genes parentales, pero no presente entre la progenie funcional, es en consecuencia el resultado de origen ancestral común.

La covariación entre un conjunto de variantes de la biblioteca puede ser evaluada y se visualiza mediante la alineación de las secuencias y la eliminación de los residuos que se conservan a lo largo de la alineación. La información mutua entre cada par de residuos diferentes se representa en una matriz de dos dimensiones. Cada fila / columna representa una de las posiciones de los residuos que varían de una proteína y cada celda de la matriz representa un posible par de residuos. Una célula de llenado de la matriz corresponde a los residuos altamente covariantes. Cada secuencia parental ha evolucionado de forma independiente a través de la selección natural y su distribución fillogenética está muy agrupadaa. Viendo cada par de residuos de los genes parentlaes identifica muchos pares de residuos que covarían. La distribución de la información mutua se normalizó al tener una media de 0 y una varianza de 1. La covariación aquí se define como pares de residuos con información mutua mayor que 2 desviaciones para que esa alineación.

Después de confeccionar la biblioteca sintética, pero antes de la exposición de las variantes para cualquier presión selectiva, se aislan las variantes. Estas variantes no apantalladas se caracterizan por la covariación en la misma forma que los genes parentales. En la mayoría de los casos, la distribución de los residuos que varían es uniforme, existiendo todos los residuos que varían en relación con todos los otros residuos variables. En la medida en que haya covariación, esa covariación no es el resultado de las limitaciones funcionales (es decir, las variantes no han estado expuestas a la selección). Esto, en efecto, es un control de la cuestión de si la covariación es el resultado de las limitaciones funcionales. Después del barajado sintético y selección para la función, pares de residuos covariados que se identifican son el resultado de las limitaciones funcionales. La covariación encontrada entre los genes parentales y no entre las variantes de la biblioteca funcionalmente activas también podría reflejar una presión selectiva para los efectos indirectos sobre el organismo. Los efectos indirectos podrían potencialmente ser cualquier rasgo, como el secuestro de los cofactores o localización celular, etc., que no está específicamente relacionada con los criterios de selección del ensayo de selección.

1. ANÁLISIS DE INFORMACIÓN MUTUA

En una alineación de proteínas, la medida de entropía para cada posición en la alineación indica el grado de variabilidad y de la preferencia para cada aminoácido. La siguiente ecuación se utiliza para cuantificar entropía de sitio (Shannon, C. E. La teoría de la comunicación matemática. 1963. MD Comput 14, 306-17. (1997)).

$$\mathbf{I_i = \sum_k P(A^k_i)\ log\ P(A^k_i)} \qquad \mathbf{(1)}$$

Donde la suma es sobre todos los aminoácidos k $\{A^k_i\}$ ocurriendo en la posición i en la alineación. P (A^k_i) es la probabilidad del aminoácido k en la posición i. Igualmente, la covarianza entre los aminoácidos se puede medir mediante el uso de la información mutua contenida entre pares de sitios.

$$\mathbf{MI_{ij} = \sum_k \sum_l P(A^k_i\ y\ A^l_j)\ log\ \frac{P(A^k_i\ y\ A^l_j)}{P(A^k_i)\ P(A^l_j)}}$$

La doble suma se extiende a todos los posibles pares de aminoácidos $\{A^k_i\}$ y $\{A^l_j\}$ en las posiciones i y j respectivamente. P (A^k_i) es la probabilidad combinada de aminoácido k en la posición i y P $(A^k_i$ y $A^l_j)$ se la probabilidad combinada de aminoácido k en la posición i y aminoácido l en la posición j.

Los valores de MI se normalizan para cada grupo de variantes para tener la misma media de 0,0 y la desviación estándar de 1,0. El grado de co-variación entre cualquier par de residuos se identifica por la desviación de la MI para el par dado del contenido esperado de información mutua.

REIVINDICACIONES

1. Un método implementado por ordenador para predecir si las secuencias de polipéptidos objetivos hipotéticos se tienen o pueden tener al menos una propiedad funcional deseada, comprendiendo el método:

(A) La identificación de uno o más motivos comunes a dos o más miembros de una población de secuencia de polipéptidos variantes usando un programa de ordenador, en el que las variantes de secuencia de polipéptido son variantes en el orden y la identidad de residuos de aminoácidos, en el que el motivo es un patrón de aminoácidos, y en el que un subconjunto de la población de secuencia de polipéptidos variantes comprende al menos una propiedad funcional deseada y un subconjunto de la población de secuencia de polipéptidos variantes carece de al menos una propiedad funcional deseada, para producir un conjunto de datos con motivos que comprende los motivos identificados en cada uno de los subgrupos de la población de variantes de secuencias de polipéptido;

(B) La correlación de al menos un motivo a partir de los datos con motivos establecidos con el que al menos una propiedad funcional deseada para producir una función de puntuación de motivo, puntuando al menos un motivo de acuerdo con una frecuencia de ocurrencia en cada subconjunto de la población de secuencia de polipéptidos variantes o de acuerdo con a una frecuencia de ausencia de cada subconjunto de la población de la secuencia de polipéptido variantes , en el que la función de puntuación de motivo es capaz de predecir si las secuencias de polipéptidos objetivos hipotéticos se tienen o pueden tener la al menos una propiedad funcional deseada; y

(C) La anotación de secuencias de polipéptidos objetivos hipotéticos, utilizando la función de puntuación de motivo para predecir si las secuencias hipotéticas de polipéptido de objetivo se tienen o pueden tener al menos una propiedad funcional deseada.

2. El método de la reivindicación 1, que comprende además la síntesis de un polipéptido que corresponde a una secuencia hipotética de polipéptido de objetivo.

3. El método de la reivindicación 1, que comprende además la anotación de secuencias hipotéticas con el fin de encontrar una o más secuencias de polipéptidos.

4. El método de la reivindicación 3, que comprende además la síntesis de uno o más secuencias de polipéptidos.

5. El método de la reivindicación 4, que comprende además el sometimiento de uno o más polipéptidos sintetizados a una o más operaciones de evolución artificial.

6. El método de la reivindicación 1, que comprende además la generación de una búsqueda libre basada en la puntuación de al menos una secuencia de polipéptido de objetivo, utilizando la función de puntuación de motivo.

7. El método de la reivindicación 1, en el que la población de variantes de secuencia de polipéptido comprende uno o más grupos de variantes de cadena de caracteres de polipéptidos que incluyen variantes positivas de cadena de caracteres de polipéptido, variantes negativas de cadena de caracteres, polipéptido y / o variantes de cadena de caracteres de polipéptido parental.

8. El método de la reivindicación 1, que comprende además la producción de la población de variantes de secuencia de polipéptido, utilizando una o más técnicas de evolución artificial, en el que al menos una operación de una o más técnicas de evolución artificial se lleva a cabo *in silico*.

9. El método de la reivindicación 1, que comprende además la producción de al menos un árbol de clasificación para organizar motivos del conjunto de datos con motivos.

10. El método de la reivindicación 1, que comprende además la realización de un Análisis de Componente Principal (PCA) de los datos con motivos establecidos para identificar una o más combinaciones de motivos que son predictivas de al menos una propiedad funcional deseada.

11. El método de la reivindicación 10, en el que el análisis de componentes principales (PCA) comprende un análisis parcial de mínimos cuadrados.

12. El método de la reivindicación 10, en el que una o más combinaciones de motivos identificas se utilizan para refinar la función de puntuación de motivo.

13. Un sistema para predecir si las secuencias hipotéticas de polipéptidos de objetivo tendrán o es probable que tengan al menos una propiedad funcional deseada, que comprende:

(A) al menos un ordenador que comprende una base de datos capaz de almacenar secuencias; y

(B) el software del sistema que comprende una o más instrucciones lógicas para:

(i) la identificación de uno o más motivos comunes a dos o más miembros de una población de variantes de secuencia de polipéptidos, donde las variantes de secuencia de polipéptido son variantes en el orden y la identidad de residuos de aminoácidos, donde el motivo es un patrón de aminoácidos, y donde un subconjunto de la población de variantes de secuencia de polipéptido comprende la al menos una propiedad funcional deseada y un subconjunto de la población de variantes de secuencia de polipéptidos carece de al menos una propiedad funcional deseada, para producir un conjunto de datos con motivos que comprende los motivos identificados en cada uno de los grupos de la población de variantes de secuencias de polipéptido;

(ii) la correlación de al menos un motivo a partir de los datos con motivos establecidos con al menos una propiedad funcional deseada para producir una función de puntuación de motivo, por la puntuación de al menos un motivo de acuerdo con una frecuencia de ocurrencia en cada subconjunto de la población de secuencia de variantes de polipéptidos o de acuerdo con a una frecuencia de ausencia de cada subconjunto de la población de la secuencia de variantes de polipéptido, donde la función de puntuación de motivo es capaz de predecir si las secuencias de polipéptidos de objetivo hipotéticos se tienen o pueden tener al menos una propiedad funcional deseada; y

(iii) la puntuación de las secuencias hipotéticas de polipéptidos de objetivo, utilizando la función de puntuación de motivo para predecir si las secuencias de polipéptidos de objetivo hipotéticas se tienen o pueden tener al menos una propiedad funcional deseada; y

(iv) la sintetización opcional de polipéptido que corresponde a una secuencia hipotética de polipéptido de objetivo.

14. Un producto de programa de ordenador para la predicción de si las secuencias polipeptídicas de objetivo hipotéticas se tienen o pueden tener al menos una propiedad funcional deseada, que comprende un medio legible por ordenador que tiene una o más instrucciones lógicas para:

(A) la identificación de uno o más motivos comunes a dos o más miembros de una población de variantes de secuencia de polipéptidos, en la que las variantes de secuencia de polipéptido son variantes en el orden y la identidad de residuos de aminoácidos, en el que el motivo es un patrón de aminoácidos, y en el que un subconjunto de la población de variantes de secuencia de polipéptido comprende al menos una propiedad funcional deseada y un subconjunto de la población de variantes de secuencia de polipéptido carece de la al menos una propiedad funcional deseada para producir un conjunto de datos con motivos que comprende los motivos identificados en cada uno de los subgrupos de la población de variantes de secuencias de polipéptido;

(B) la correlación de al menos un motivo a partir de los datos con motivos establecidos con al menos una propiedad funcional deseada para producir una función de puntuación de motivo, por puntuación por lo menos un motivo de acuerdo con una frecuencia de ocurrencia en cada subconjunto de la población de la secuencia de variantes de polipéptido o de acuerdo con una frecuencia de ausencia de cada subconjunto de la población de variantes de la secuencia de polipéptido, donde la función de puntuación de motivo es capaz de predecir si las secuencias de polipéptidos de objetivo hipotéticas se tienen o pueden tener la al menos una propiedad funcional deseada; y

(C) la puntuación de secuencias de polipéptidos de objetivo hipotéticas, utilizando la función de puntuación de motivo para predecir si las secuencias de polipéptidos de objetivo hipotéticas se tienen o pueden tener al menos una propiedad funcional deseada; y

(D) la sintetización opcional de un polipéptido que corresponde a una secuencia hipotética de polipéptido de objetivo.

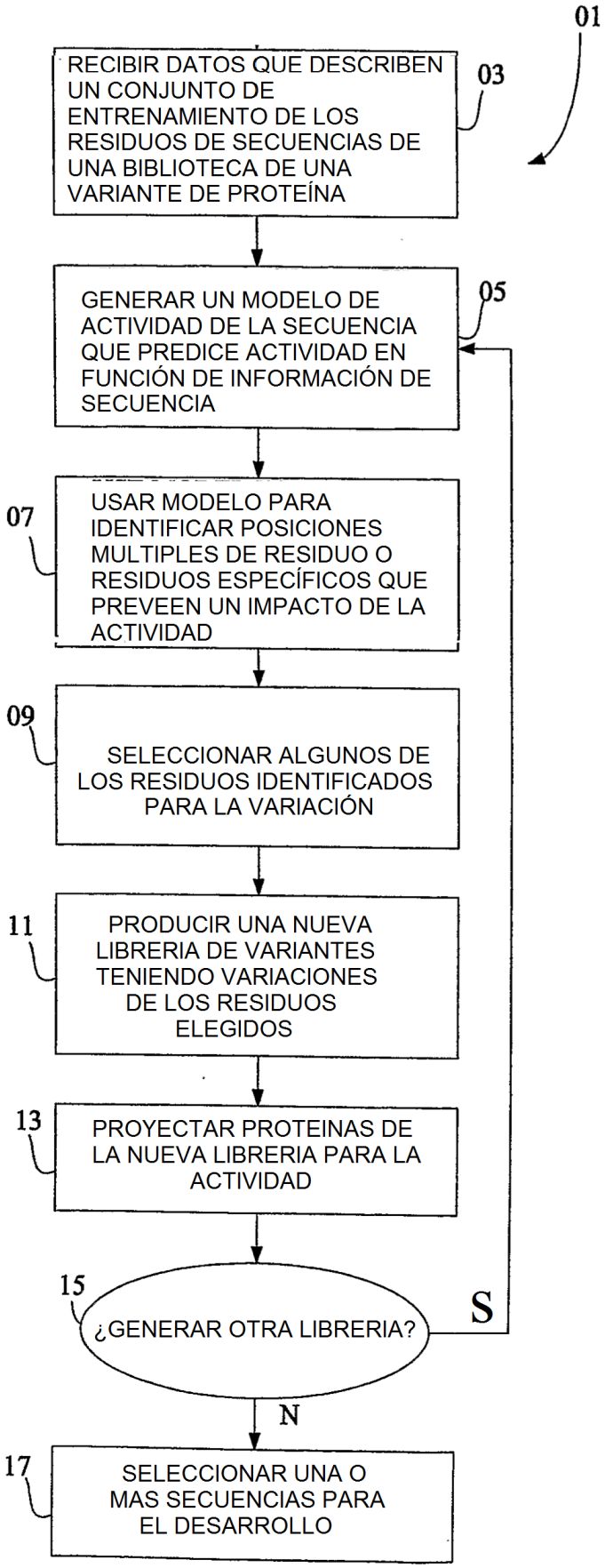
01
RECIBIR DATOS QUE DESCRIBEN UN CONJUNTO DE ENTRENAMIENTO DE LOS RESIDUOS DE SECUENCIAS DE UNA BIBLIOTECA DE UNA VARIANTE DE PROTEÍNA
03
GENERAR UN MODELO DE ACTIVIDAD DE LA SECUENCIA QUE PREDICE ACTIVIDAD EN FUNCIÓN DE INFORMACIÓN DE SECUENCIA
05
USAR MODELO PARA IDENTIFICAR POSICIONES MULTIPLES DE RESIDUO O RESIDUOS ESPECÍFICOS QUE PREVEEN UN IMPACTO DE LA ACTIVIDAD
07
SELECCIONAR ALGUNOS DE LOS RESIDUOS IDENTIFICADOS PARA LA VARIACIÓN
09
PRODUCIR UNA NUEVA LIBRERIA DE VARIANTES TENIENDO VARIACIONES DE LOS RESIDUOS ELEGIDOS
11
PROYECTAR PROTEINAS DE LA NUEVA LIBRERIA PARA LA ACTIVIDAD
13
¿GENERAR OTRA LIBRERIA?
15
S
N
SELECCIONAR UNA O MAS SECUENCIAS PARA EL DESARROLLO
17

Fig. 1

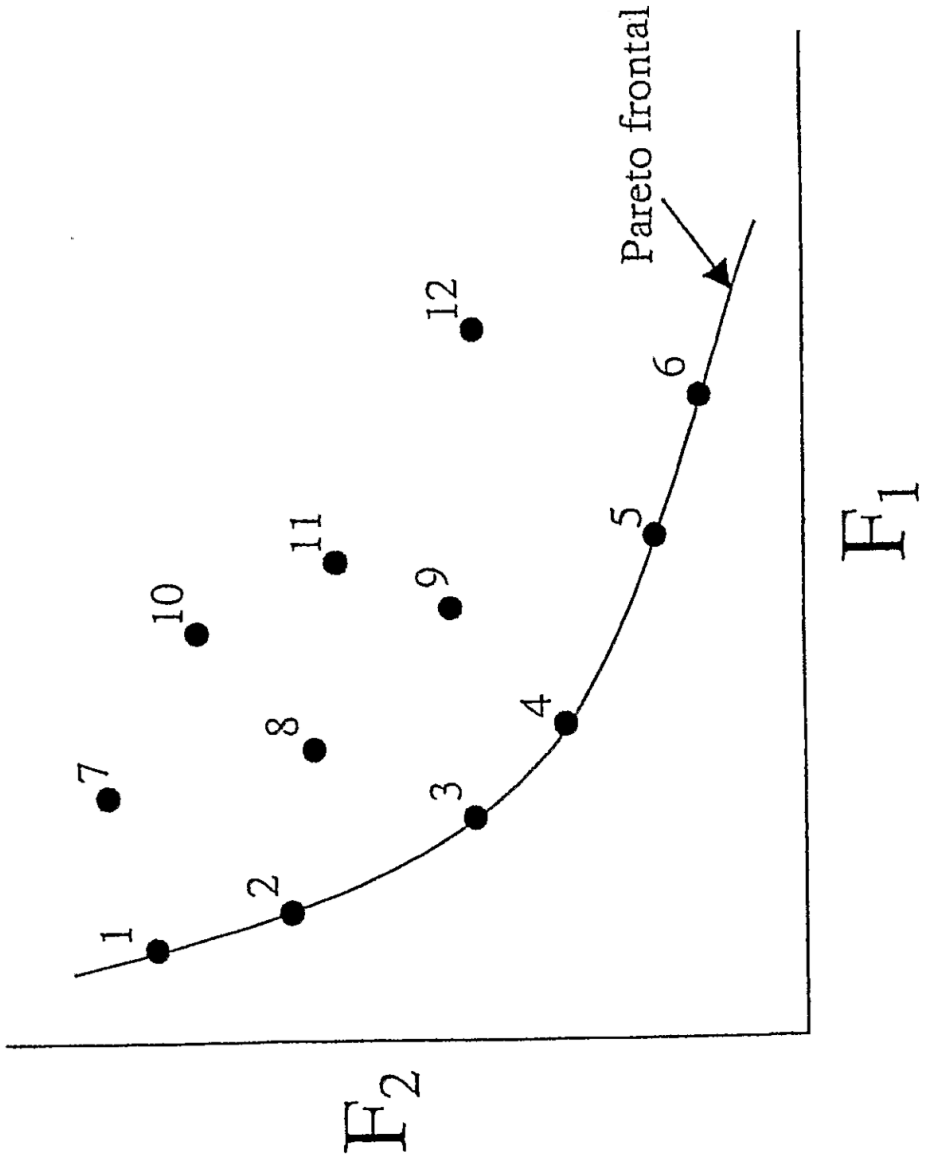

Fig. 2

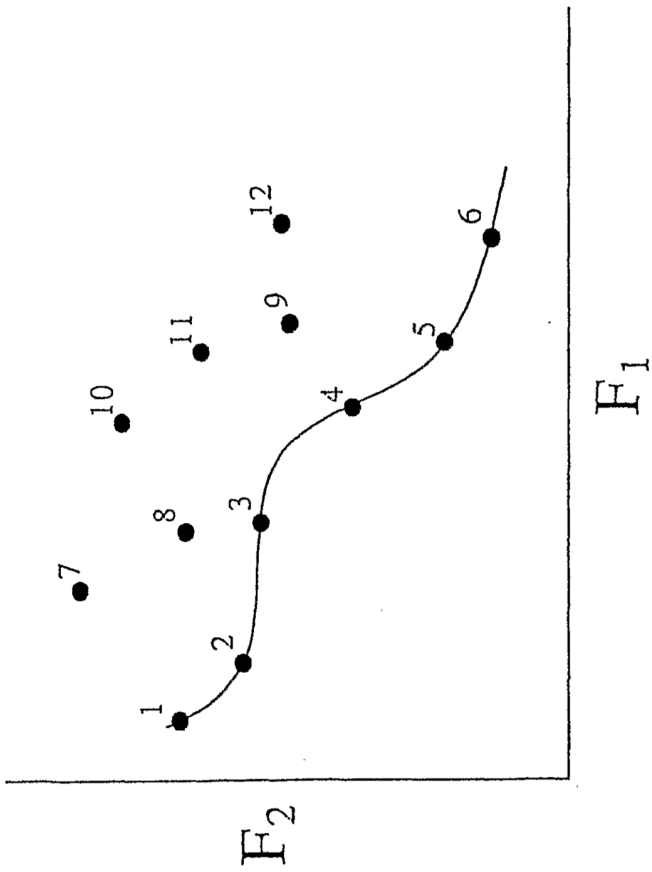

Fig. 3

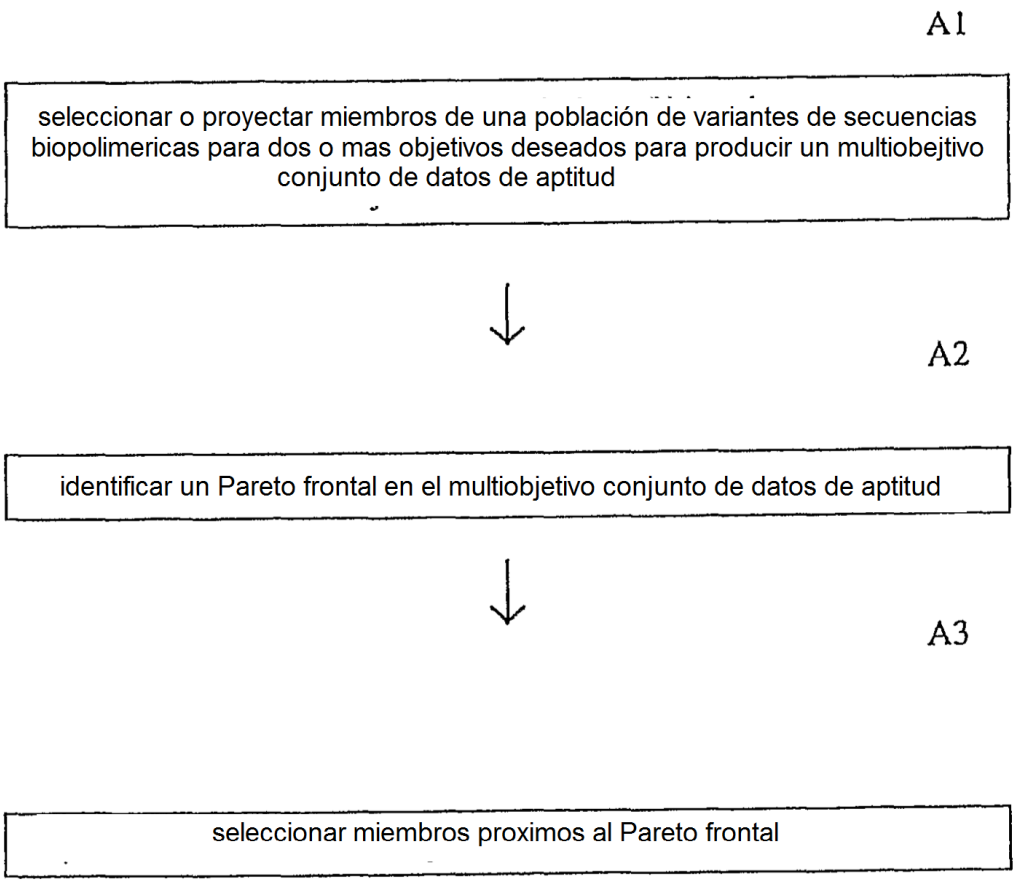
A1
seleccionar o proyectar miembros de una población de variantes de secuencias biopolimericas para dos o mas objetivos deseados para producir un multiobejtivo conjunto de datos de aptitud
A2
identificar un Pareto frontal en el multiobjetivo conjunto de datos de aptitud
A3
seleccionar miembros proximos al Pareto frontal

Fig. 4

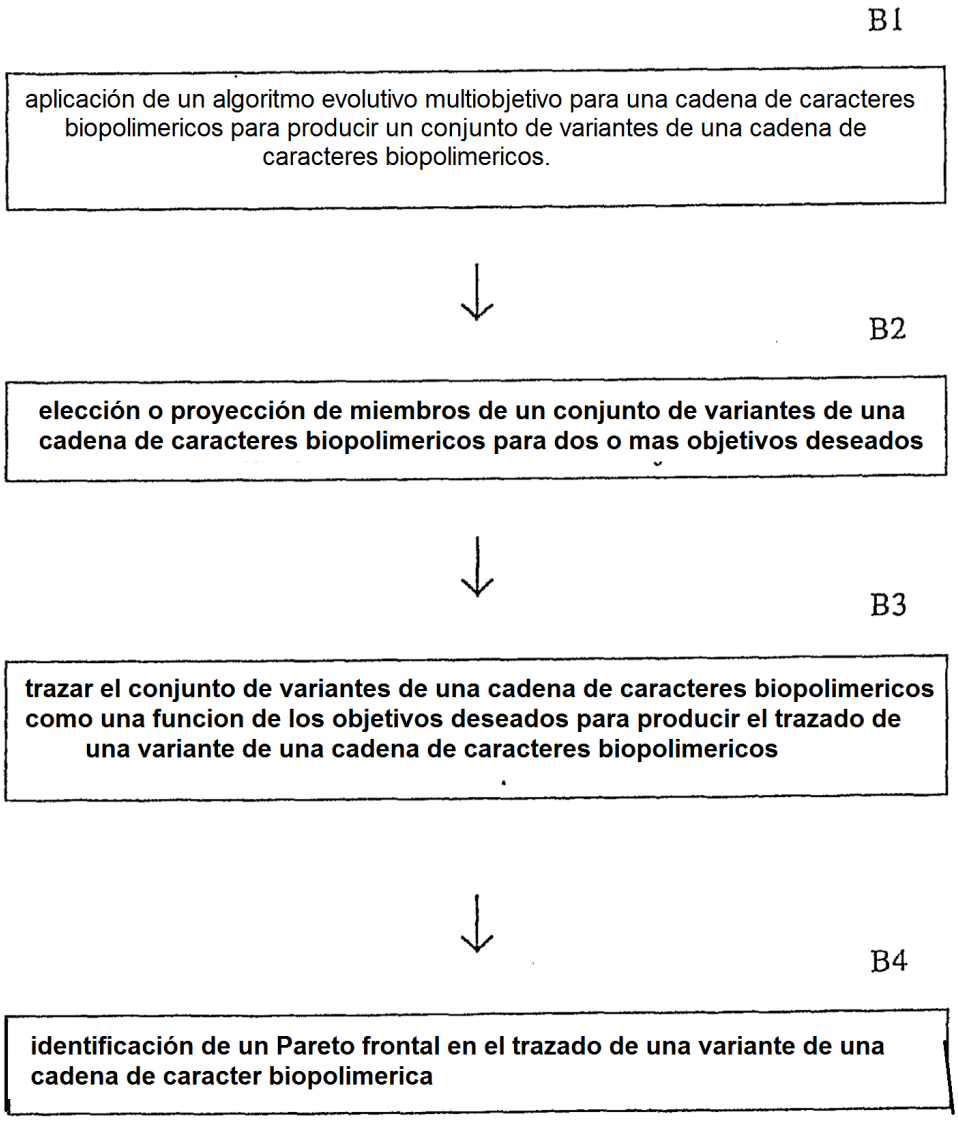
B1
aplicación de un algoritmo evolutivo multiobjetivo para una cadena de caracteres biopolimericos para producir un conjunto de variantes de una cadena de caracteres biopolimericos.
B2
elección o proyección de miembros de un conjunto de variantes de una cadena de caracteres biopolimericos para dos o mas objetivos deseados
B3
trazar el conjunto de variantes de una cadena de caracteres biopolimericos como una funcion de los objetivos deseados para producir el trazado de una variante de una cadena de caracteres biopolimericos
B4
identificación de un Pareto frontal en el trazado de una variante de una cadena de caracter biopolimerica

Fig. 5

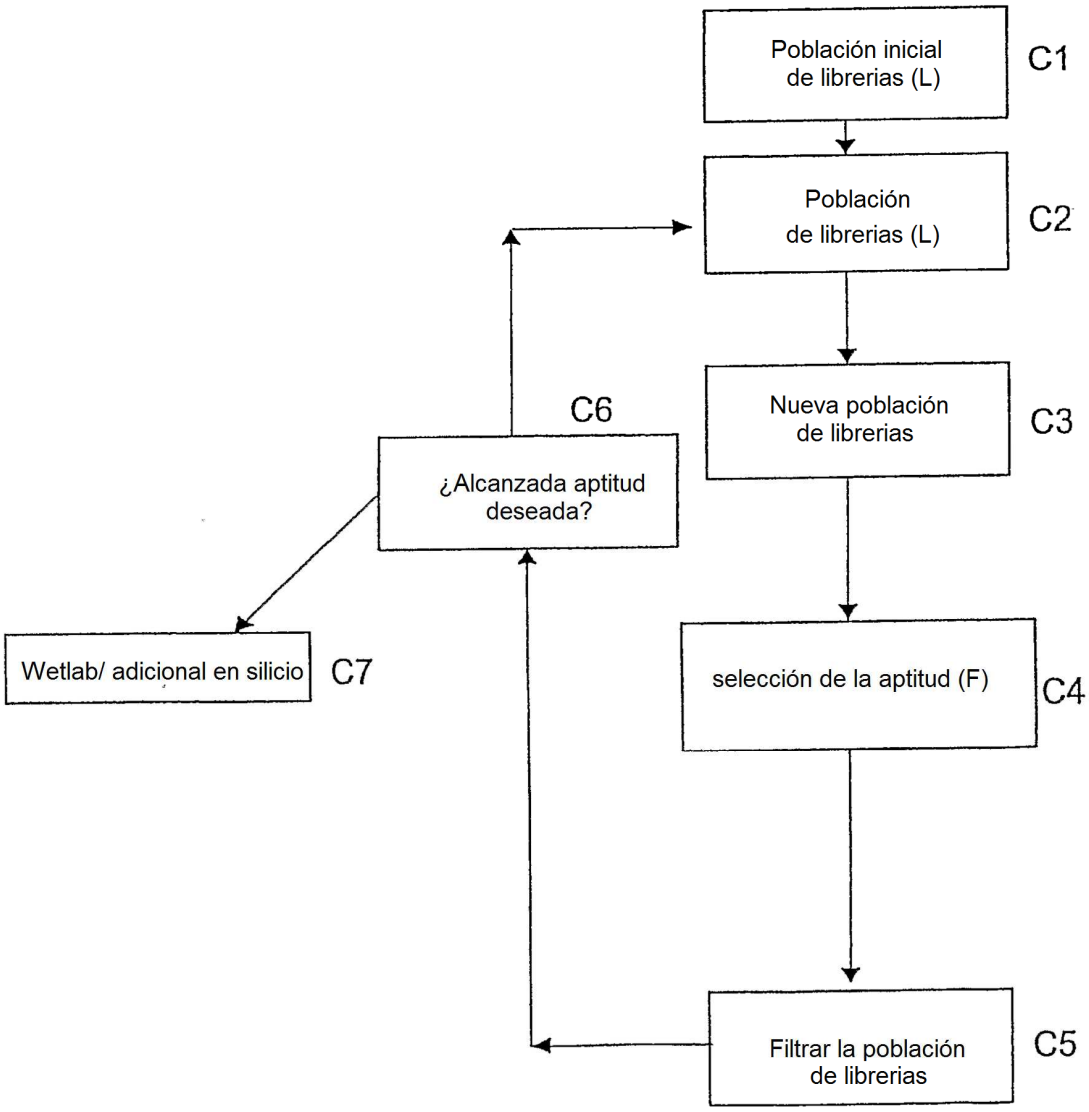
Población inicial de librerias (L)
C1
Población de librerias (L)
C2
Nueva población de librerias
C3
C6
¿Alcanzada aptitud deseada?
Wetlab/ adicional en silicio
C7
selección de la aptitud (F)
C4
Filtrar la población de librerias
C5

Fig. 6

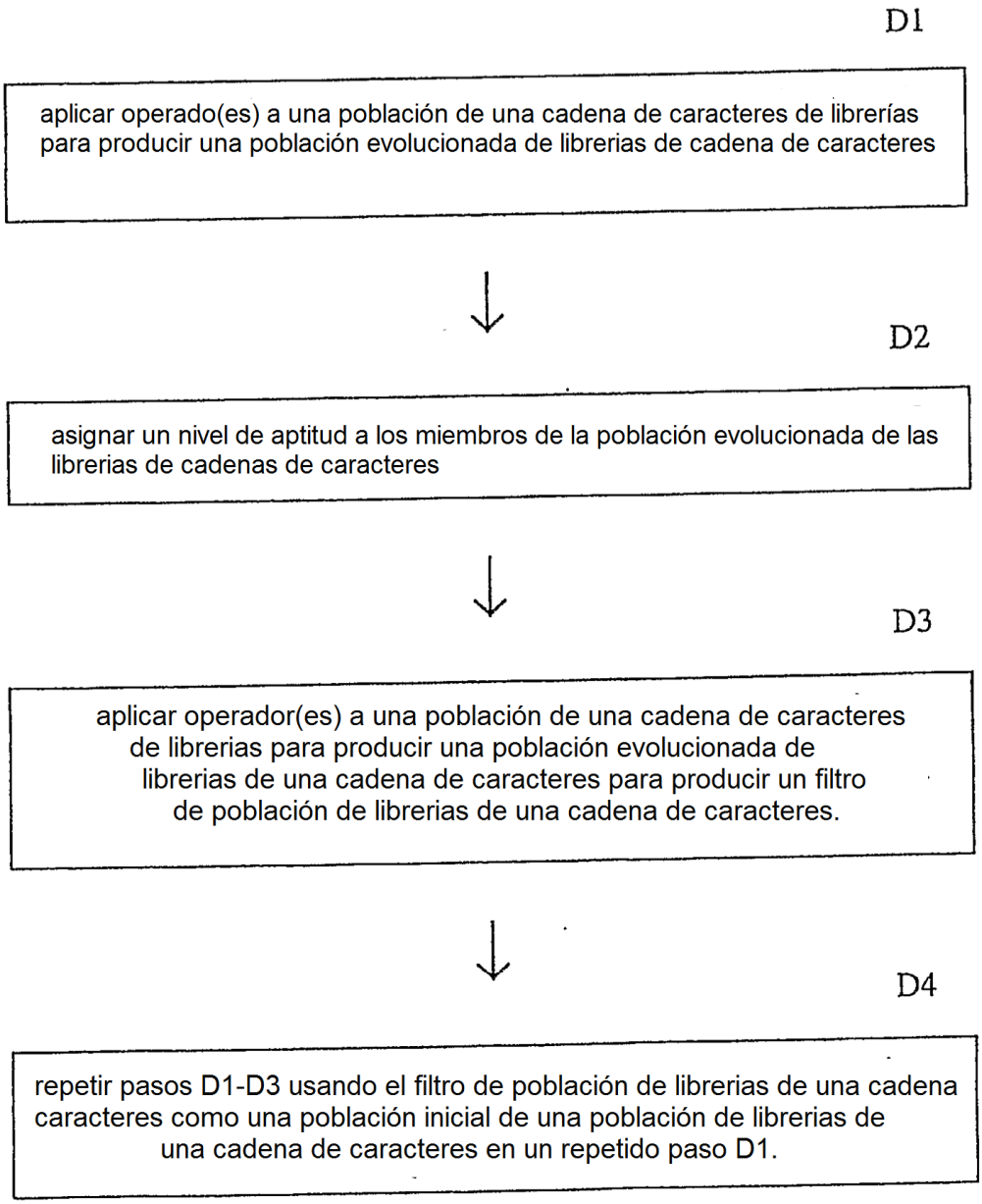
D1
aplicar operado(es) a una población de una cadena de caracteres de librerías para producir una población evolucionada de librerias de cadena de caracteres
D2
asignar un nivel de aptitud a los miembros de la población evolucionada de las librerias de cadenas de caracteres
D3
aplicar operador(es) a una población de una cadena de caracteres de librerias para producir una población evolucionada de librerias de una cadena de caracteres para producir un filtro de población de librerias de una cadena de caracteres.
D4
repetir pasos D1-D3 usando el filtro de población de librerias de una cadena caracteres como una población inicial de una población de librerias de una cadena de caracteres en un repetido paso D1.

Fig. 7

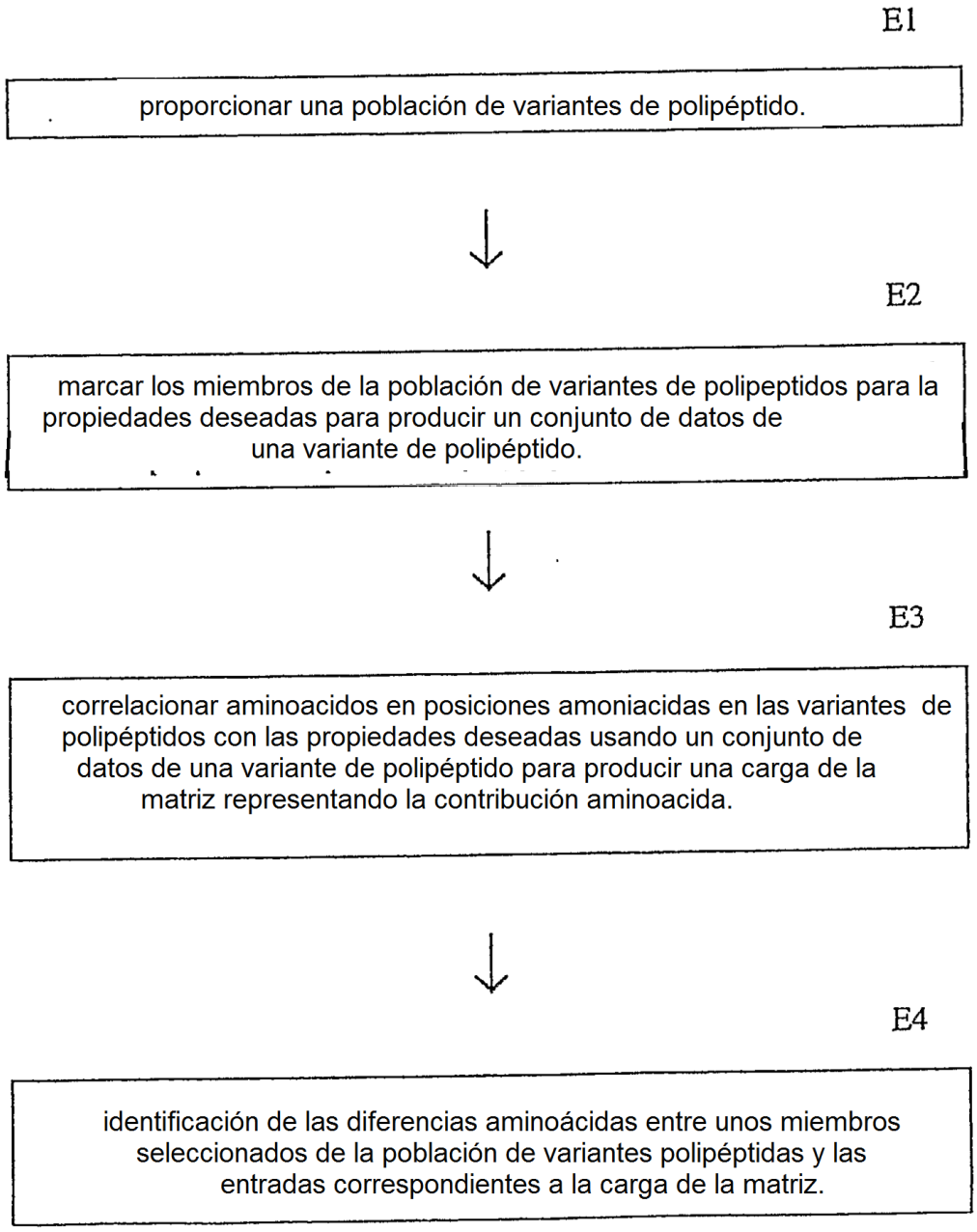
E1
proporcionar una población de variantes de polipéptido.
E2
marcar los miembros de la población de variantes de polipeptidos para la propiedades deseadas para producir un conjunto de datos de una variante de polipéptido.
E3
correlacionar aminoacidos en posiciones amoniacidas en las variantes de polipéptidos con las propiedades deseadas usando un conjunto de datos de una variante de polipéptido para producir una carga de la matriz representando la contribución aminoacida.
E4
identificación de las diferencias aminoácidas entre unos miembros seleccionados de la población de variantes polipéptidas y las entradas correspondientes a la carga de la matriz.

Fig. 8

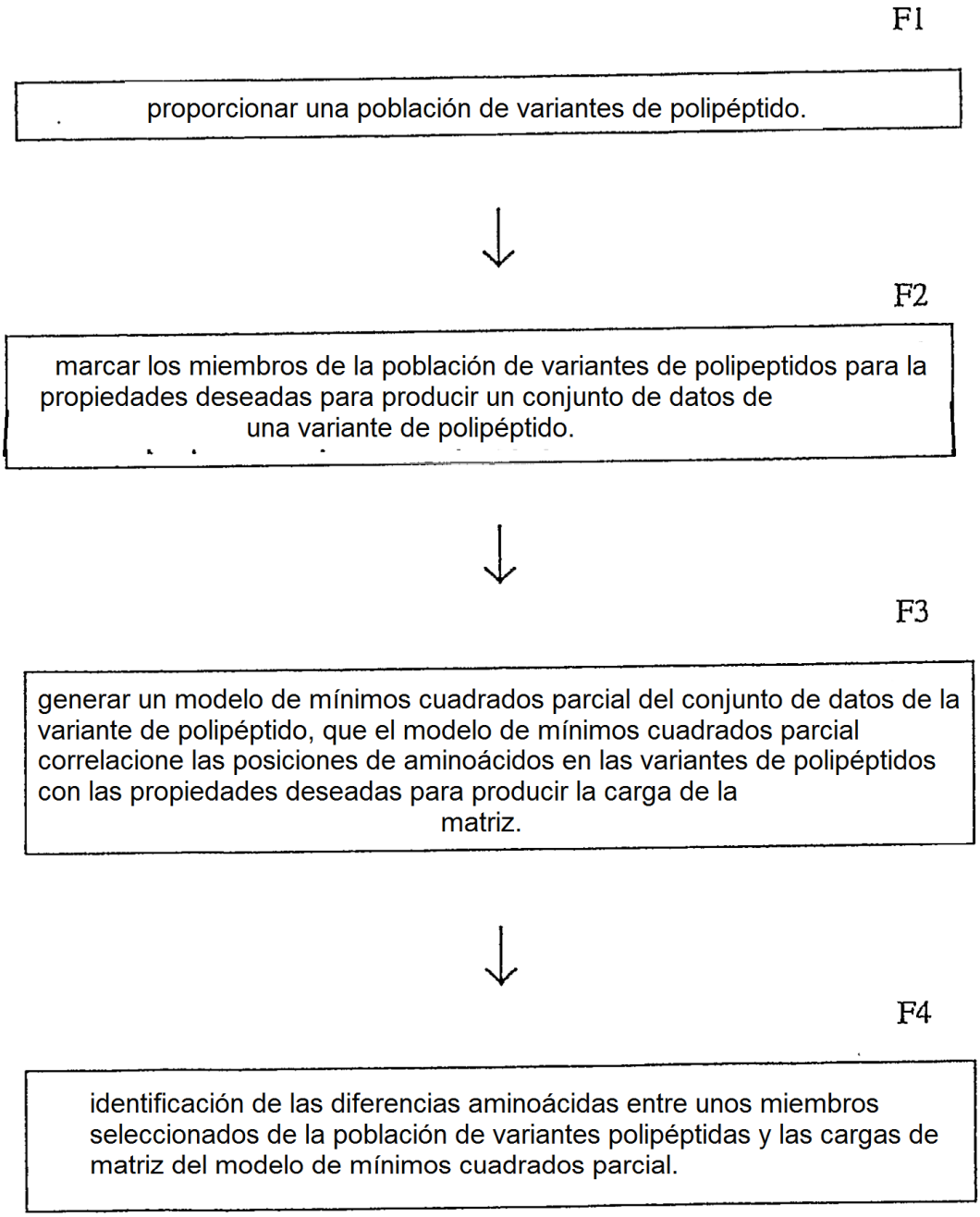
F1
proporcionar una población de variantes de polipéptido.
F2
marcar los miembros de la población de variantes de polipeptidos para la propiedades deseadas para producir un conjunto de datos de una variante de polipéptido.
F3
generar un modelo de mínimos cuadrados parcial del conjunto de datos de la variante de polipéptido, que el modelo de mínimos cuadrados parcial correlacione las posiciones de aminoácidos en las variantes de polipéptidos con las propiedades deseadas para producir la carga de la matriz.
F4
identificación de las diferencias aminoácidas entre unos miembros seleccionados de la población de variantes polipéptidas y las cargas de matriz del modelo de mínimos cuadrados parcial.

Fig. 9

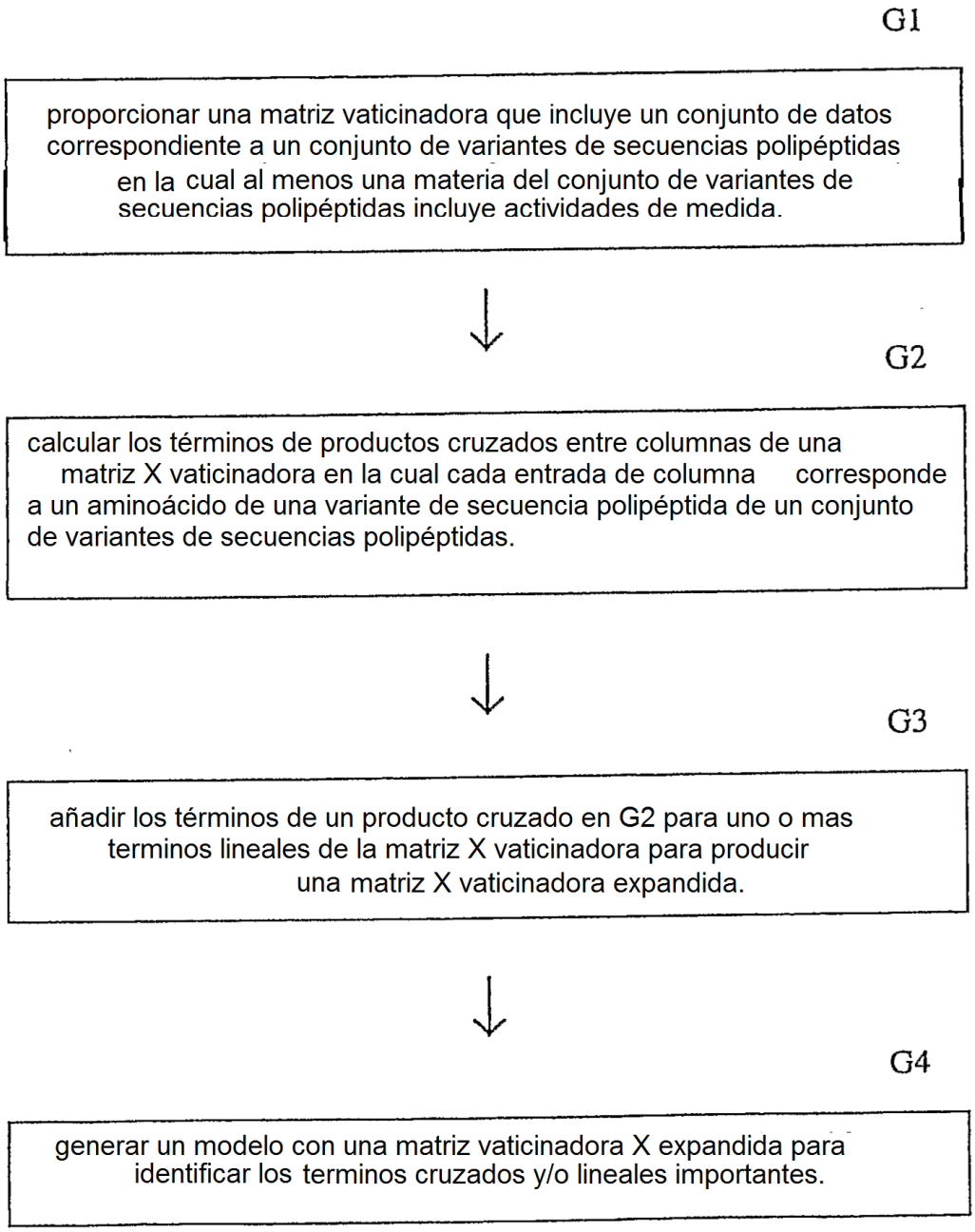
G1
proporcionar una matriz vaticinadora que incluye un conjunto de datos correspondiente a un conjunto de variantes de secuencias polipéptidas en la cual al menos una materia del conjunto de variantes de secuencias polipéptidas incluye actividades de medida.
G2
calcular los términos de productos cruzados entre columnas de una matriz X vaticinadora en la cual cada entrada de columna corresponde a un aminoácido de una variante de secuencia polipéptida de un conjunto de variantes de secuencias polipéptidas.
G3
añadir los términos de un producto cruzado en G2 para uno o mas terminos lineales de la matriz X vaticinadora para producir una matriz X vaticinadora expandida.
G4
generar un modelo con una matriz vaticinadora X expandida para identificar los terminos cruzados y/o lineales importantes.

Fig. 10

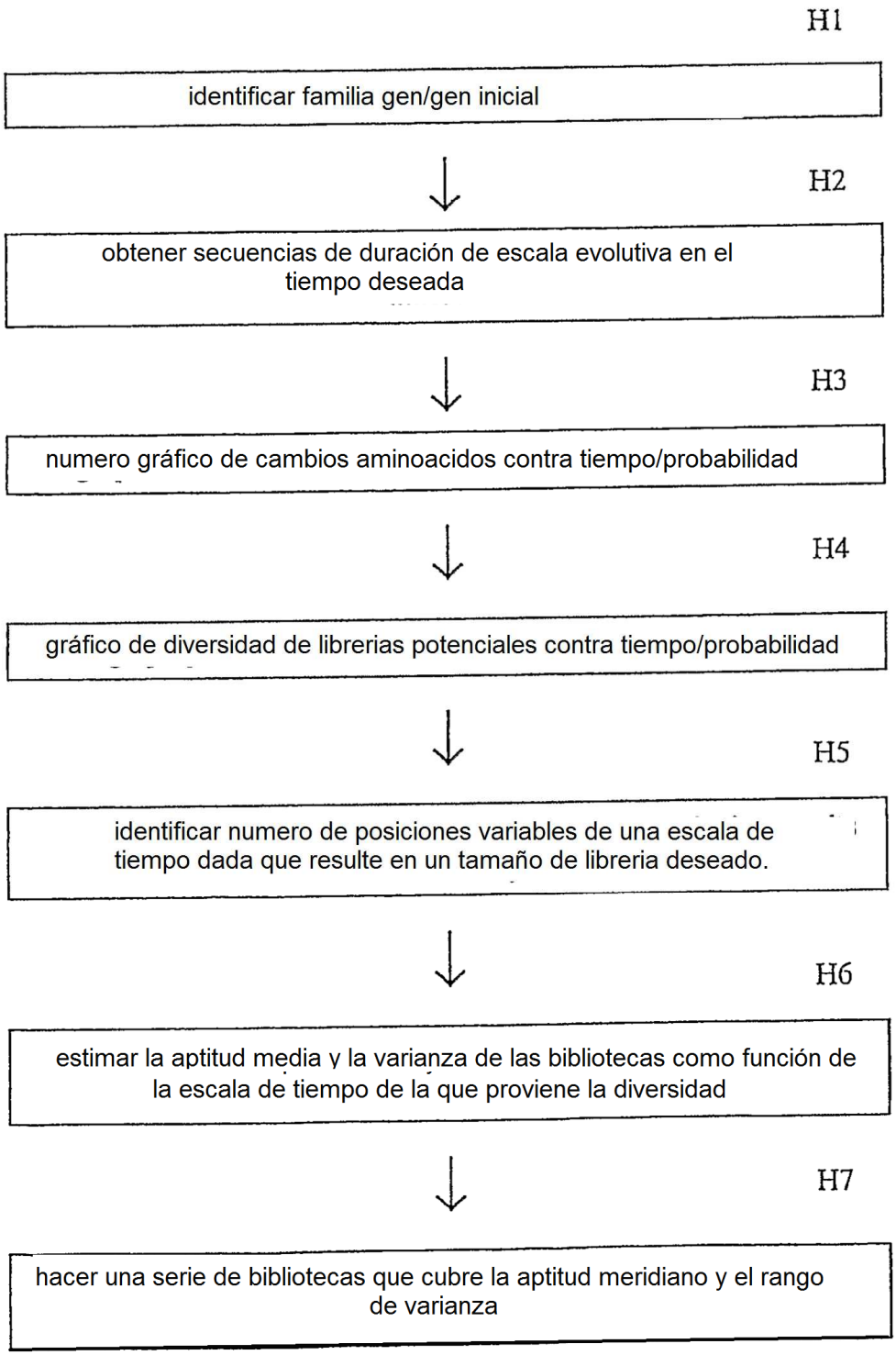
H1
identificar familia gen/gen inicial
H2
obtener secuencias de duración de escala evolutiva en el tiempo deseada
H3
numero gráfico de cambios aminoacidos contra tiempo/probabilidad
H4
gráfico de diversidad de librerias potenciales contra tiempo/probabilidad
H5
identificar numero de posiciones variables de una escala de tiempo dada que resulte en un tamaño de libreria deseado.
H6
estimar la aptitud media y la varianza de las bibliotecas como función de la escala de tiempo de la que proviene la diversidad
H7
hacer una serie de bibliotecas que cubre la aptitud meridiano y el rango de varianza

Fig. 11

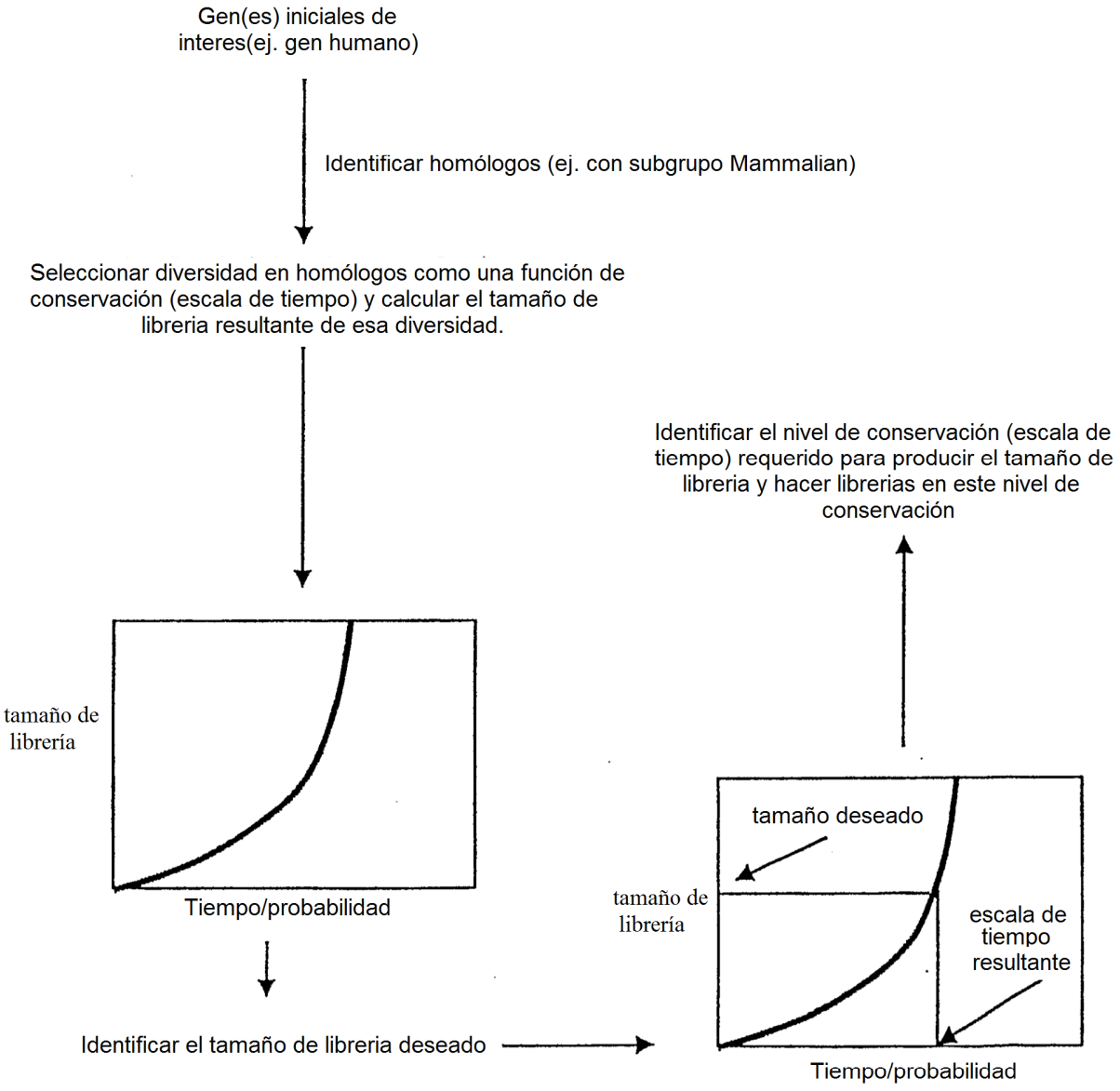
Gen(es) iniciales de interes(ej. gen humano)
Identificar homólogos (ej. con subgrupo Mammalian)
Seleccionar diversidad en homólogos como una función de conservación (escala de tiempo) y calcular el tamaño de libreria resultante de esa diversidad.
tamaño de librería
Tiempo/probabilidad
Identificar el tamaño de libreria deseado
Identificar el nivel de conservación (escala de tiempo) requerido para producir el tamaño de libreria y hacer librerias en este nivel de conservación
tamaño deseado
tamaño de librería
escala de tiempo resultante
Tiempo/probabilidad

Fig. 12

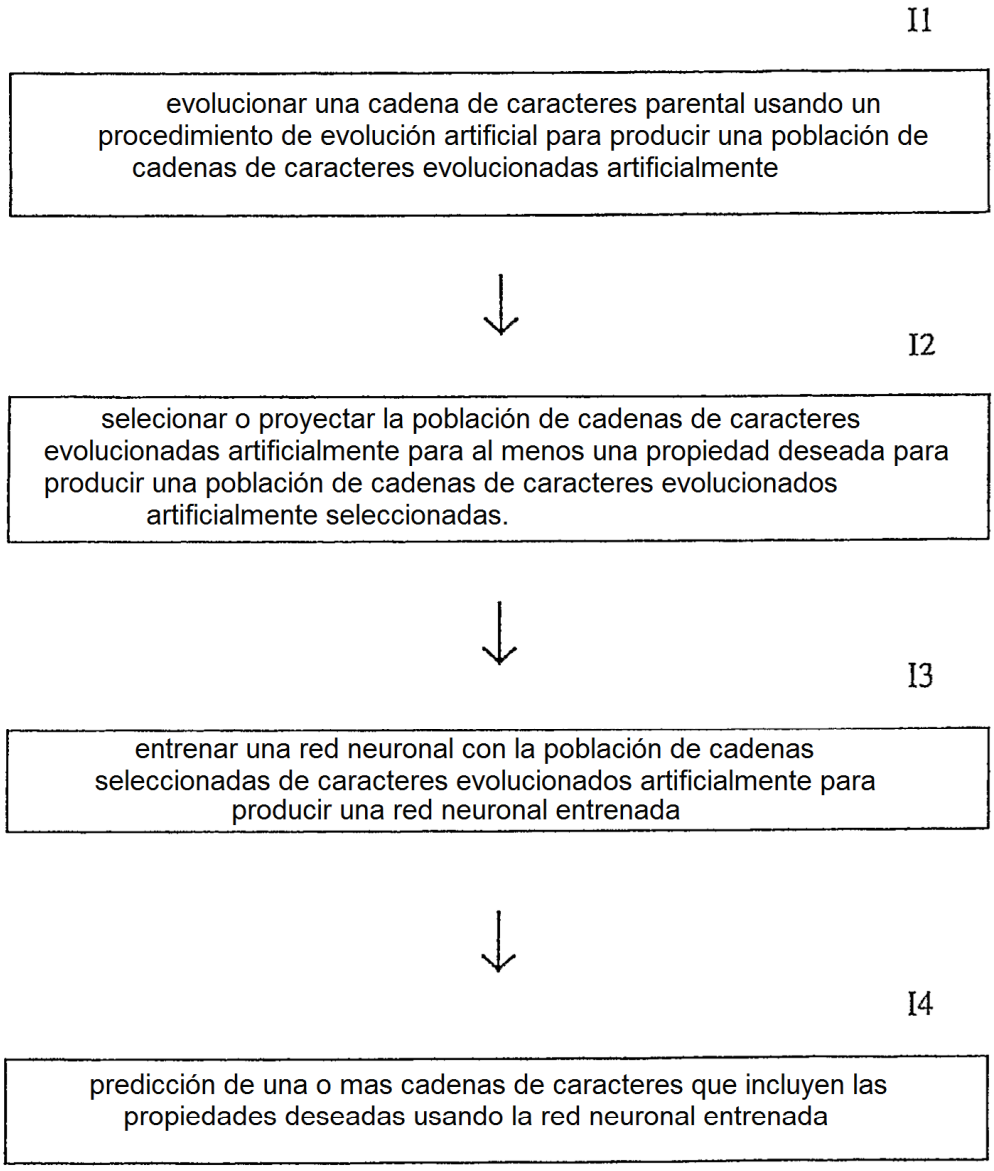
I1
evolucionar una cadena de caracteres parental usando un procedimiento de evolución artificial para producir una población de cadenas de caracteres evolucionadas artificialmente
I2
selecionar o proyectar la población de cadenas de caracteres evolucionadas artificialmente para al menos una propiedad deseada para producir una población de cadenas de caracteres evolucionados artificialmente seleccionadas.
I3
entrenar una red neuronal con la población de cadenas seleccionadas de caracteres evolucionados artificialmente para producir una red neuronal entrenada
I4
predicción de una o mas cadenas de caracteres que incluyen las propiedades deseadas usando la red neuronal entrenada

Fig. 13

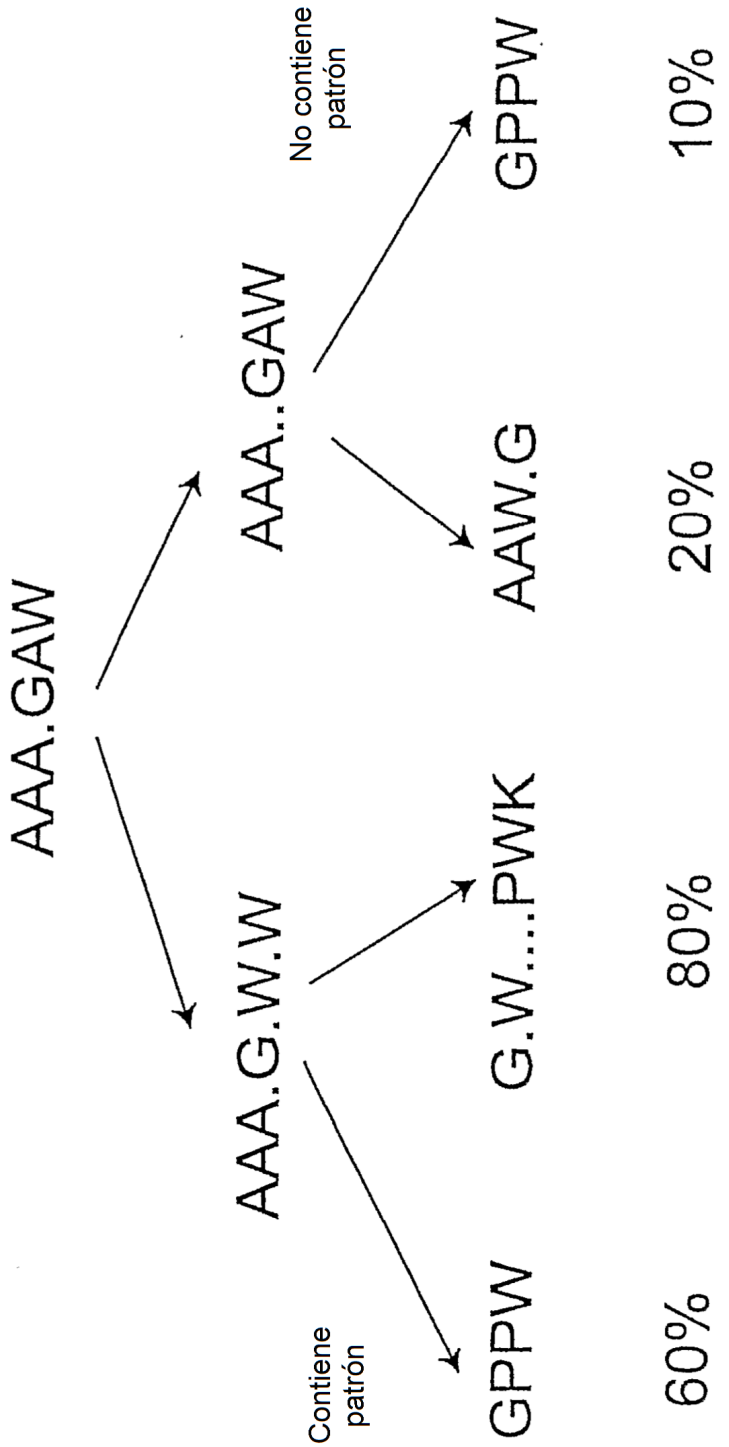
AAA.GAW
AAA.G.W.W
AAA..GAW
Contiene patrón
No contiene patrón
GPPW
G.W....PWK
AAW.G
GPPW
60%
80%
20%
10%

Fig. 14

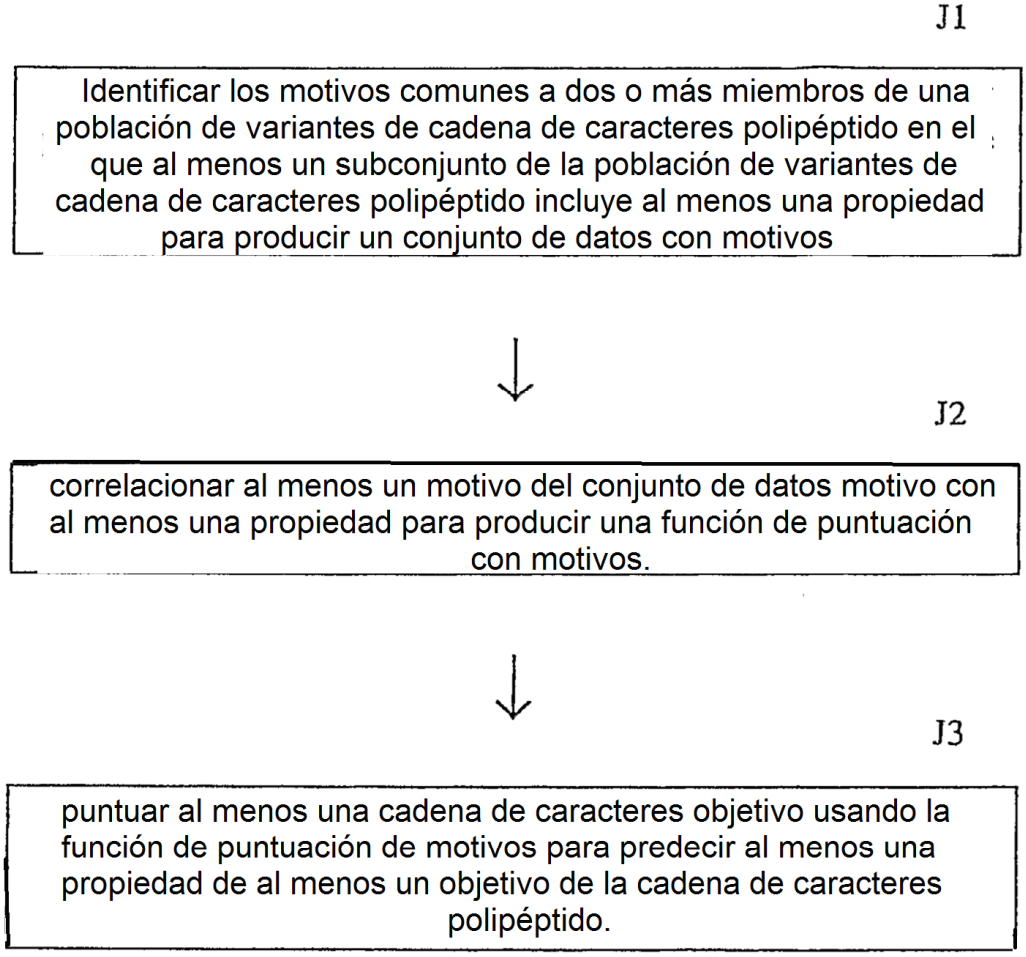
J1
Identificar los motivos comunes a dos o más miembros de una población de variantes de cadena de caracteres polipéptido en el que al menos un subconjunto de la población de variantes de cadena de caracteres polipéptido incluye al menos una propiedad para producir un conjunto de datos con motivos
J2
correlacionar al menos un motivo del conjunto de datos motivo con al menos una propiedad para producir una función de puntuación con motivos.
J3
puntuar al menos una cadena de caracteres objetivo usando la función de puntuación de motivos para predecir al menos una propiedad de al menos un objetivo de la cadena de caracteres polipéptido.

Fig. 15

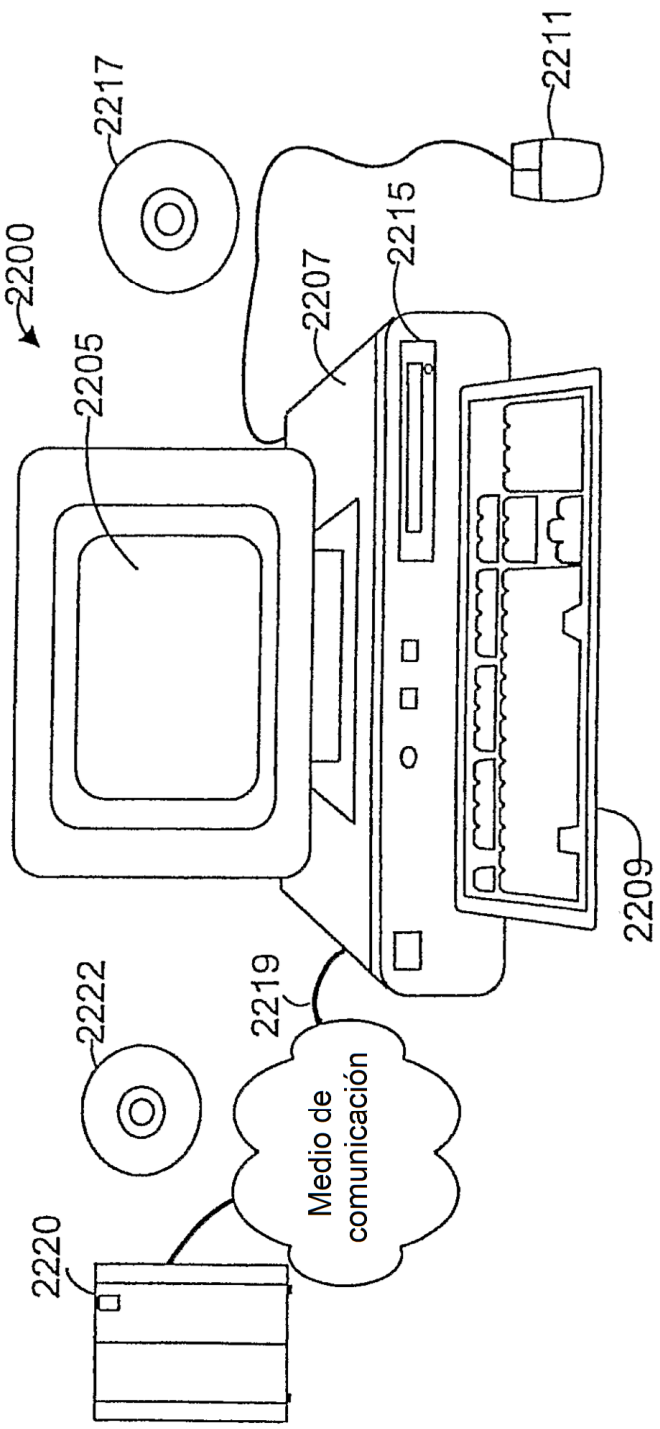
2200
2205
2207
2209
2211
2215
2217
2219
2220
2222
Medio de comunicación

Fig. 16