(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 774 965**

(51) Int. CI.:

G16B 15/30 (2009.01)
G16B 35/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **26.09.2014 PCT/US2014/057900**

(87) Fecha y número de publicación internacional: **02.04.2015 WO15048573**

(96) Fecha de presentación y número de la solicitud europea: **26.09.2014 E 14786396 (3)**

(97) Fecha y número de publicación de la concesión europea: **01.01.2020 EP 3049979**

(54) Título: **Modelado predictivo a base de estructura**

(30) Prioridad:

27.09.2013 US 201361883919 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
23.07.2020

(73) Titular/es:

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

(72) Inventor/es:

SARMIENTO, RUSSELL JAVINIAR;
BASKERVILLE, DONALD SCOTT y
ZHANG, XIYUN

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Modelado predictivo a base de estructura

ANTECEDENTES

[0001] Desde hace tiempo se sabe que el diseño de proteínas es una tarea difícil si no es por otra razón que la explosión combinatoria de posibles moléculas que constituyen el espacio de secuencia de búsqueda. El espacio de secuencia de las proteínas es inmenso y es imposible de explorar exhaustivamente utilizando métodos actualmente conocidos en la técnica, que a menudo están limitados por el tiempo y el costo requeridos para identificar polipéptidos útiles. Parte del problema surge de la gran cantidad de variantes de polipéptidos que deben secuenciarse, seleccionarse y analizarse. Los métodos de evolución dirigida aumentan la eficiencia para perfeccionar las biomoléculas candidatas que tienen propiedades ventajosas. Hoy, la evolución dirigida de las proteínas está dominada por varios cribado de alto rendimiento y formatos de recombinación, a menudo realizados de forma iterativa.

[**0002]** También se han propuesto diversas técnicas computacionales para explorar el espacio de secuencias-actividad. Relativamente hablando, estas técnicas están en su infancia y todavía se necesitan avances significativos. En consecuencia, los nuevos métodos para mejorar la eficiencia de la detección, secuenciación y ensayo de biomoléculas candidatas son altamente deseables.

[**0003]** El documento US 2006/136184 A1 describe métodos, sistemas informáticos y productos de programas informáticos para la ingeniería de biopolímeros.

[**0004]** Chaparro-Riggers et al. (Biotechnol J. 2007 Feb; 2 (2): 180-91) describe nuevas estrategias en el campo del motor de proteínas basado en datos ering, elegir residuos de aminoácidos particulares para variar a fin de aumentar las posibilidades de obtener una proteína variante con la propiedad deseada.

[**0005]** Estados Unidos 2009/0118130 A1 describe composiciones y métodos para el cribado de una pluralidad de variantes de polipéptido.

[**0006]** Fox y col. (Nat Biotechnol. 2007 Mar; 25 (3): 338-44) describen un enfoque de evolución dirigida que aumenta la evolución dirigida basada en recombinación al incorporar una estrategia para el análisis estadístico de las relaciones de actividad de la secuencia de proteínas (ProSAR).

SUMARIO

[0007] La invención proporciona un procedimiento implementado por ordenador de la realización de la evolución dirigida, comprendiendo el método:

(a) recibir un conjunto de datos sin filtrar que tiene información de mediciones físicas de las moléculas, en el que el conjunto de datos sin filtrar comprende la siguiente información para cada uno de una pluralidad de biomoléculas variantes: (i) actividad de la biomolécula variante en un ligando en un sitio de unión de la biomolécula variante, (ii) una secuencia de la biomolécula variante, en la que la secuencia es una secuencia de ácido nucleico o una secuencia de proteína, y (iii) uno o más parámetros geométricos que caracterizan la geometría del ligando en el sitio de unión de la biomolécula variante;

(b) filtrar el conjunto de datos sin filtrar para producir un subconjunto de datos filtrados, eliminando información para una o más de las biomoléculas variantes, en donde el filtrado elimina al menos uno de los parámetros geométricos del conjunto de datos sin filtrar y/o elimina del conjunto de datos sin filtrar ciertas biomoléculas variantes que tienen valores de parámetros geométricos fuera de los rangos definidos, y en donde el filtrado comprende modelos de actividad de secuencia de entrenamiento con una pluralidad de subconjuntos de datos seleccionados y prueba de la capacidad de los modelos de actividad de secuencia entrenados con la pluralidad de subconjuntos de datos seleccionados para predecir la actividad de una biomolécula variante en el ligando en el sitio de unión de la biomolécula variante en función de variables independientes y, por lo tanto, identifica un subconjunto de datos filtrados que proporciona un modelo de actividad de secuencia con mayor capacidad para predecir la actividad de una biomolécula variante en el ligando en el sitio de unión de la biomolécula variante en función de variables independientes que un modelo de actividad de secuencia entrenado con el conjunto de datos sin filtrar, en el que la secuencia de la biomolécula variante y los parámetros geométricos filtrados que caracterizan la geometría del ligando en el sitio de unión de la biomolécula variante son variables independientes en los modelos de actividad de secuencia entrenados con la pluralidad de subconjuntos de datos seleccionados, cada subconjunto de datos seleccionado tiene información para al menos uno de los parámetros geométricos y/o ciertas biomoléculas variantes que tienen valores de parámetros geométricos fuera de los rangos definidos eliminados del conjunto de datos sin filtrar de (a); y

(c) aplicar un modelo de actividad de secuencia entrenado usando el subconjunto de datos filtrados para

identificar una o más variantes de biomolécula nuevas predichas para tener actividad que cumpla con uno o más criterios, en donde cada una de las una o más variantes de biomolécula nuevas tienen una secuencia de ácido nucleico o proteína que difiere de las secuencias de las variantes de biomoléculas que proporcionan información para el conjunto de datos sin filtrar de (a).

[**0008]** La invención proporciona además un producto de programa de ordenador que comprende uno o más medios de almacenamiento no transitorios legibles por ordenador que tiene instrucciones ejecutables almacenadas por ordenador que, cuando es ejecutado por uno o más procesadores de un sistema informático, causa que el sistema de ordenador implemente un método de la invención.

[0009] La invención proporciona además un sistema de ordenador, que comprende:

uno o más procesadores;

memoria del sistema; y

uno o más medios de almacenamiento legibles por computadora que han almacenado en instrucciones ejecutables por computadora que, cuando son ejecutadas por uno o más procesadores, hacen que el sistema informático implemente un método para llevar a cabo la evolución dirigida de acuerdo con un método de la invención.

[**0010]** La presente descripción se refiere a los campos de la biología molecular, evolución molecular, bioinformática y sistemas digitales.

[**0011]** Los métodos de la presente divulgación tienen utilidad en la optimización de proteínas para uso industrial y terapéutico. Los métodos y sistemas son especialmente útiles para diseñar y desarrollar enzimas que tienen propiedades o actividades beneficiosas.

[**0012]** Ciertos aspectos de la presente divulgación se refieren a métodos para desarrollar proteínas que tienen propiedades beneficiosas y/o guían programas de evolución dirigida. La descripción presenta métodos para identificar biomoléculas con propiedades deseadas (o que son más adecuadas para la evolución dirigida hacia tales propiedades) a partir de bibliotecas complejas de biomoléculas o conjuntos de dichas bibliotecas. Algunas realizaciones de la presente descripción proporcionan métodos para construir un modelo de actividad de secuencia con referencia a datos estructurales, dicho modelo puede usarse para guiar la evolución dirigida de proteínas que tienen propiedades beneficiosas. Algunas realizaciones usan algoritmo(s) genético(s) y datos estructurales para filtrar datos no informativos. Algunas realizaciones usan máquinas de vectores de soporte para entrenar el modelo de actividad de secuencia. Los métodos de filtrado y entrenamiento pueden generar un modelo de actividad de secuencia que tiene mayor poder predictivo que los métodos de modelado convencionales.

[**0013]** Algunas realizaciones de la descripción proporcionan métodos para la realización de evolución dirigida. En algunas realizaciones, el método se implementa usando un sistema informático que incluye uno o más procesadores y memoria del sistema. El método incluye: (a) recibir un conjunto de datos que tiene información de mediciones físicas de moléculas, en el que el conjunto de datos incluye la siguiente información para cada una de una pluralidad de biomoléculas variantes: (i) actividad de la biomolécula variante en un ligando en un enlace sitio de la biomolécula variante, (ii) una secuencia de la biomolécula variante, y (iii) uno o más parámetros geométricos que caracterizan la geometría del ligando en el sitio de unión; (b) filtrar el conjunto de datos para producir un subconjunto de datos filtrados mediante la eliminación de información para una o más de las biomoléculas variantes, en donde el filtrado comprende probar el poder predictivo de los modelos de actividad de secuencia entrenados con una pluralidad de subconjuntos de datos seleccionados, cada subconjunto de datos seleccionado tiene información para un conjunto particular de biomoléculas variantes eliminadas del conjunto de datos de (a); y (c) entrenar un modelo de actividad de secuencia mejorado utilizando el subconjunto de datos filtrados. En algunas realizaciones, la información para cada una de la pluralidad de biomoléculas variantes también incluye (iv) una energía de interacción que caracteriza la interacción del ligando en el sitio de unión. En algunas realizaciones, las biomoléculas variantes son enzimas.

[**0014]** En algunas realizaciones, el modelo de actividad de la secuencia mejorada se obtiene mediante una máquina de vectores de soporte, una regresión lineal múltiple, una regresión de componentes principales, una regresión parcial de mínimos cuadrados, o una red neural.

[**0015]** En algunas realizaciones, la filtración de la serie de datos implica la eliminación de al menos uno de los parámetros geométricos de la serie de datos. En algunas realizaciones, el filtrado del conjunto de datos se realiza con un algoritmo genético. En algunas realizaciones, el algoritmo genético varía los umbrales para eliminar la información asociada con los parámetros geométricos para una o más de las biomoléculas variantes.

[**0016]** En algunas realizaciones, el método para la evolución dirigida implica además la aplicación de la secuencia de la mejora de modelo de actividad para identificar uno o más nueva biomolécula variantes predicho por el modelo de actividad secuencia mejorada para tener actividad cumplimiento de ciertos criterios. Cada una de las nuevas variantes

de biomoléculas tiene una secuencia que difiere de las secuencias de las variantes de biomoléculas que proporcionan información para el conjunto de datos de (a). En algunas realizaciones, la aplicación del modelo de actividad de secuencia mejorada para identificar una o más variantes nuevas de biomolécula implica realizar un algoritmo genético en donde se evalúan nuevas variantes potenciales de biomolécula usando el modelo de actividad de secuencia mejorada como una función de aptitud.

[**0017]** En algunas realizaciones, el método para la evolución dirigida adicional implica el ensayo de las nuevas variantes biomoleculares para la actividad. En algunas realizaciones, el método también implica la medición de la actividad de las biomoléculas variantes por un ensayo *in vitro*.

[**0018]** En algunas realizaciones, el procedimiento implica además la producción de un modelo estructural para cada uno de las nuevas variantes biomoleculares. El método también utiliza los modelos estructurales para generar parámetros geométricos para los sitios de unión de las nuevas variantes de biomoléculas. Los parámetros geométricos caracterizan la geometría del ligando en los sitios de unión de las nuevas variantes de biomoléculas. En algunas realizaciones, el método implica además recibir modelos estructurales de variantes de biomoléculas y determinar el uno o más parámetros geométricos usando los modelos estructurales. En algunas realizaciones, los modelos estructurales son modelos de homología. En algunas realizaciones, los modelos de homología se preparan usando detalles de medición estructural física de biomoléculas. Los detalles de la medición estructural física de las biomoléculas pueden incluir posiciones tridimensionales de átomos obtenidos por RMN o cristalografía de rayos X.

[**0019]** En algunas realizaciones, el procedimiento implica además el uso de un cargador de muelle para determinar uno o más parámetros geométricos. En algunas realizaciones, el método también usa un acoplador para determinar la energía de interacción. En algunas realizaciones, las variantes biomoleculares procesadas son una pluralidad de enzimas. En algunas realizaciones, la actividad de la biomolécula variante en un ligando es la actividad de una enzima en un sustrato. En algunas realizaciones, la actividad de una enzima en un sustrato incluye una o más características de una conversión catalítica del sustrato por la enzima.

[**0020]** En algunas realizaciones, el método para la evolución dirigida también implica el uso del modelo de actividad de la secuencia mejorada para identificar una o más biomoléculas que tienen actividad deseada. En algunas realizaciones, el método incluye además sintetizar las biomoléculas que tienen la actividad deseada.

[**0021]** En algunas realizaciones, también se proporcionan los productos de programas de ordenador y sistemas informáticos de aplicación de los métodos para la evolución dirigida de biomoléculas.

[**0022]** Estas y otras características se presentará a continuación con referencia a los dibujos asociados.

Breve descripción de los dibujos

[**0023]**

La figura 1A es un diagrama de flujo que representa un flujo de trabajo de evolución dirigida de acuerdo con algunas realizaciones de la divulgación.

La figura 1B es un diagrama de flujo que ilustra un enfoque para filtrar datos sin procesar de acuerdo con algunas realizaciones de la divulgación.

La figura 1C es un diagrama de flujo que presenta un proceso de filtrado de datos de acuerdo con algunas realizaciones, en donde la etapa de selección de características no se realiza o se combina con la etapa de selección de distancia.

La figura 2 muestra tres representaciones tabulares de un conjunto de datos de actividad de secuencia para ilustrar un ejemplo de filtrado de datos de acuerdo con algunas realizaciones de la divulgación actual.

La figura 3A es un diagrama de flujo que muestra un algoritmo genético para filtrar datos sin procesar para eliminar uno o más tipos de energía y/o tipos de geometría de acuerdo con algunas realizaciones de la divulgación.

La figura 3B es un diagrama de flujo que muestra un algoritmo genético para filtrar datos sin procesar para eliminar datos de variantes que tienen valores de energía y/o valores de geometría fuera de los rangos definidos de acuerdo con algunas realizaciones de la divulgación.

La figura 3C es un diagrama de flujo que muestra un algoritmo genético para identificar nuevas variantes de biomoléculas usando un modelo de actividad de secuencia de alto poder predictivo de acuerdo con algunas realizaciones de la divulgación.

La figura 4 muestra un dispositivo digital ejemplar que puede implementarse de acuerdo con algunas

realizaciones.

DESCRIPCIÓN DETALLADA

[**0024]** Se describen métodos para desarrollar modelos de actividad de secuencia con referencia a datos estructurales. Los modelos de actividad de secuencia pueden usarse para guiar la evolución dirigida de proteínas que tienen propiedades beneficiosas. Algunas realizaciones pueden ayudar a explorar un gran espacio de secuencia y perfeccionar rápidamente moléculas de propiedades beneficiosas. Los materiales y/o recursos también se pueden guardar en los procesos para encontrar o desarrollar proteínas de las propiedades deseadas. Algunas realizaciones son especialmente útiles para diseñar y desarrollar enzimas que tienen actividad y/o selectividad deseadas para reacciones catalíticas que implican sustratos particulares.

I. DEFINICIONES

[0025] A menos que se defina lo contrario en el presente documento, todos los términos técnicos y científicos usados en este documento tienen el mismo significado que se entiende comúnmente por un experto ordinario en la técnica. Varios diccionarios científicos que incluyen los términos incluidos aquí son bien conocidos y están disponibles para los expertos en la materia. Cualquier método y material similar o equivalente a los descritos en este documento encuentran uso en la práctica de las realizaciones descritas en este documento.

[**0026]** Los términos definidos inmediatamente a continuación se entenderán más completamente por referencia a la memoria descriptiva como un todo. Las definiciones se proporcionan con el propósito de describir realizaciones particulares solamente y ayudar a comprender los conceptos complejos descritos en esta especificación. No tienen la intención de limitar el alcance total de la divulgación. Específicamente, debe entenderse que esta descripción no se limita a las secuencias, composiciones, algoritmos, sistemas particulares, metodologías, protocolos y/o reactivos descritos en el presente documento, ya que estos pueden variar, dependiendo del contexto en que los usan los expertos en la materia.

[**0027]** Como se usa en esta memoria descriptiva y las reivindicaciones adjuntas, las formas singulares "un", "una", "el" y “ella” incluyen los referentes plurales a menos que el contenido y el contexto claramente dicte otra cosa. Así, por ejemplo, la referencia a "un dispositivo" incluye una combinación de dos o más de tales dispositivos, y similares. A menos que se indique lo contrario, una conjunción "o" está destinada a utilizarse en su sentido correcto como operador lógico booleano, abarcando tanto la selección de características en la alternativa (A o B, donde la selección de A es mutuamente excluyente de B) como la selección de características en conjunto (A o B, donde se seleccionan A y B).

[**0028]** Máquinas de vectores de soporte (SVMs) son herramientas de aprendizaje de máquina con algoritmos de aprendizaje asociados para la clasificación y análisis de regresión. El SVM básico toma un conjunto de datos de entrada y predice, para cada entrada dada, cuál de las dos clases posibles forma la salida. Dado un conjunto de ejemplos de entrenamiento, cada uno marcado como perteneciente a una de dos categorías, un algoritmo de entrenamiento SVM construye un modelo que asigna nuevos ejemplos en una categoría u otra. Un SVM es una representación de los ejemplos como puntos en el espacio, mapeados para que los ejemplos de las categorías separadas se dividan por un espacio claro que sea lo más amplio posible, que se implementa maximizando la distancia entre los puntos de datos y un hiperplano que separa dos categorías. Además de realizar una clasificación lineal, los SVM pueden realizar eficientemente una clasificación no lineal utilizando un truco del núcleo para mapear implícitamente las entradas en espacios de características de alta dimensión.

[**0029]** Cuando se usa para la optimización de los modelos de actividad de la secuencia, SVM toma como entradas conjuntos de entrenamiento de secuencias que han sido clasificados en dos o más grupos en base a la actividad. Las máquinas de vectores de soporte operan ponderando diferentes miembros de un conjunto de entrenamiento de manera diferente dependiendo de lo cerca que estén de una interfaz de hiperplano que separa a los miembros "activos" e "inactivos" del conjunto de entrenamiento. Esta técnica requiere que el científico primero decida qué miembros del grupo de entrenamiento colocarán en el grupo activo y qué miembros del grupo de entrenamiento se colocarán en el grupo inactivo. Esto se puede lograr eligiendo un valor numérico apropiado de actividad para servir como el límite entre los miembros activos e inactivos del conjunto de entrenamiento. A partir de esta clasificación, la máquina de vectores de soporte generará un vector, W, que puede proporcionar valores de coeficientes para los valores individuales de las variables independientes que definen las secuencias de los miembros del grupo activo e inactivo en el conjunto de entrenamiento. Estos coeficientes se pueden usar para "clasificar" los residuos individuales como se describe en otra parte del presente documento. La técnica intenta identificar un hiperplano que maximice la distancia entre los miembros más cercanos del conjunto de entrenamiento en lados opuestos de ese plano. En otra variación, se lleva a cabo el modelado de regresión vectorial de soporte. En este caso, la variable dependiente es un vector de valores de actividad continua. El modelo de regresión del vector de soporte generará un vector de coeficiente, W, que puede usarse para clasificar los residuos individuales.

[**0030]** Las SVM se han usado para observar grandes conjuntos de datos en muchos estudios y han sido bastante populares en el campo de microarrays de ADN. Sus fortalezas potenciales incluyen la capacidad de discriminar finamente (ponderando) qué factores separan las muestras entre sí. En la medida en que un SVM pueda determinar

con precisión qué residuos contribuyen a funcionar, puede ser una herramienta particularmente útil para clasificar los residuos de acuerdo con esta descripción. Los SVM se describen en S. Gunn (1998) "Support Vector Machines for Classification and Regressions," Technical Report, Faculty of Engineering and Applied Science, Department of Electronics and Computer Science, Universidad de Southampton.

[**0031]** Acoplador (software de acoplamiento o programa de acoplamiento): un "acoplador" es un programa informático que predice computacionalmente si un ligando se unirá o acoplará con un sitio de unión de interés en una proteína u otra molécula biológica. El proceso por el cual un ligando se acerca y finalmente se une al sitio de unión a veces se denomina "acoplamiento". El concepto de acoplamiento puede entenderse como una interacción que hace que el ligando se una con la biomolécula de tal manera que el ligando no se desaloja fácilmente. En el acoplamiento exitoso, el ligando y la biomolécula forman un complejo estable. Un ligando acoplado puede actuar como agonista o antagonista. Un acoplador puede simular y/o caracterizar el acoplamiento.

[**0032]** Los acopladores se implementan típicamente como software que puede almacenarse temporal o permanentemente en asociación con hardware tal como un procesador o procesadores. Los programas de acoplamiento disponibles comercialmente incluyen CDocker (Accelrys), DOCK (Universidad de California, San Francisco), AutoDock (Scripps Research Institute), FlexX (tripos.com), GOLD (ccdc.cam.ac.uk) y GLIDE (schrodinger.com).

[**0033]** Varios acopladores producen una puntuación de acoplamiento u otra medida de la unión entre el ligando y la biomolécula. Para algunas combinaciones de ligando-biomolécula, el programa de acoplamiento determinará que es poco probable que se produzca la unión. En tales casos, el programa de acoplamiento generará una conclusión de que el ligando no se une con la biomolécula.

[**0034]** Los acopladores pueden generar "posturas" de ligandos con respecto a los sitios de unión. Algunas de estas posturas se pueden usar para generar una puntuación de acoplamiento o de otra manera evaluar el acoplamiento. En algunas realizaciones, el acoplador permite a un usuario especificar una serie de posturas (n) para usar en la evaluación del acoplamiento. Solo las mejores posturas "n" con los mejores puntajes de acoplamiento se consideran al evaluar el acoplamiento.

[**0035]** Se puede programar un acoplador para emitir una evaluación de la probabilidad de que un ligando se acople con el sitio de unión de la biomolécula o la calidad de dicho acoplamiento, en caso de que ocurra. En un nivel, un acoplador determina si es probable que un ligando se una a un sitio de unión de biomolécula. Si la lógica del acoplador concluye que la unión no es probable o es altamente desfavorable, puede generar un resultado de "no se encontraron posturas refinadas". Esto puede ocurrir cuando todas las conformaciones que generó el programa de acoplamiento tienen choques desfavorables de van der Waals y/o repulsiones electrostáticas con el sitio de unión. En el ejemplo anterior de un procedimiento de acoplamiento, si la segunda operación no logra encontrar una postura con una energía suave menor que el umbral, el Acoplador puede dar un resultado como "no se encontraron posturas refinadas". Debido a que la energía blanda considera principalmente las interacciones no unidas, incluyendo van der Waals y las fuerzas electrostáticas, el resultado de posturas no refinadas significa que el ligando tiene choques estéricos severos y/o repulsiones electrostáticas con el receptor de biomoléculas para un número dado de posturas.

[**0036]** En ciertas realizaciones, la ventana acoplable da salida a una puntuación de acoplamiento que representa la interacción entre el ligando y el sitio de unión de biomoléculas. Los acopladores pueden calcular diversas características de la interacción ligando-biomolécula. En un ejemplo, la salida es simplemente la energía de interacción entre el ligando y la biomolécula. En otra realización, se emite una energía total. Se puede entender que la energía total es una combinación de energía de interacción ligando-biomolécula y cepa de ligando. En ciertas implementaciones, dicha energía puede calcularse utilizando un campo de fuerza como CHARMm.

[**0037]** En diversas realizaciones, los programas de Acoplador generan tales salidas considerando múltiples posturas del ligando en el sitio de unión de la biomolécula. Cada postura tendrá sus propios valores de energía asociados. En algunas realizaciones, el programa de acoplamiento clasifica las posturas y considera la energía asociada con una o más de las posturas de alto rango. En algunos casos, puede promediar las energías de ciertas posturas de alto rango o realizar un análisis estadístico de las posturas de alto rango. En otras realizaciones, simplemente elige el valor asociado con la postura mejor clasificada y genera esto como la energía resultante para el acoplamiento.

[**0038]** Una "postura" es la posición o la orientación de un ligando con respecto a un sitio de unión de una molécula biológica. En una postura, las posiciones tridimensionales de algunos o todos los átomos del ligando se especifican con respecto a algunas o todas las posiciones de los átomos en el sitio de unión. Si bien la conformación de un ligando no es su postura, debido a que la conformación no considera el sitio de unión, la conformación puede usarse para determinar una postura. En algunas realizaciones, la orientación y conformación de un ligando juntas definen una postura. En algunas realizaciones, una postura solo existe si la combinación de orientación/conformación de un ligando cumple con un nivel umbral de energía definido en el sitio de unión de referencia.

[**0039]** Varios mecanismos computacionales se pueden emplear para generar posturas para acoplamiento. Los ejemplos incluyen búsquedas torsionales sistemáticas o estocásticas sobre enlaces rotativos, simulaciones de

dinámica molecular y algoritmos genéticos para "evolucionar" nuevas conformaciones de baja energía. Estas técnicas se utilizan para modificar las representaciones computacionales del ligando y/o el sitio de unión para explorar el "espacio de postura".

[**0040]** Los acopladores evalúan posturas para determinar cómo interactúa el ligando con el sitio de unión. En algunas realizaciones, hacen esto calculando la energía de interacción basada en uno o más de los tipos de interacción mencionados anteriormente (por ejemplo, fuerzas de van der Waals). Esta información se utiliza para caracterizar el acoplamiento y, en algunos casos, producir una puntuación de acoplamiento. En algunas implementaciones, los acopladores clasifican las posturas según la puntuación de acoplamiento. En algunas implementaciones, los acopladores eliminan las posturas con puntajes de acoplamiento desfavorables.

[**0041]** En ciertas realizaciones, un sistema de evaluaciónde proteína virtual evalúa una postura para determinar si la postura es activa. Una postura se considera activa si cumple con las restricciones definidas que se sabe que son importantes para la actividad deseada bajo consideración. Como ejemplo, el sistema virtual de detección de proteínas puede determinar si una postura admite la transformación catalítica del ligando en un sitio de unión.

[**0042]** Un "ligando" es una molécula o complejo que interactúa con un sitio de unión de una biomolécula para formar un complejo estable que contiene al menos el ligando y la biomolécula. Además del ligando y la biomolécula, el complejo estable puede incluir (a veces requiere) otras entidades químicas tales como cofactores orgánicos e inorgánicos (por ejemplo, coenzimas y grupos protésicos), iones metálicos y similares. Los ligandos pueden ser agonistas o antagonistas.

[**0043]** Cuando la biomolécula es una enzima, el sitio de unión es un sitio catalítico y el ligando es un sustrato, una reacción intermedia del sustrato, o un estado de transición del sustrato. Un "intermedio de reacción" es una entidad química generada a partir del sustrato en la transformación de sustrato a producto de reacción. Un "estado de transición" de un sustrato es el sustrato en un estado correspondiente a la energía potencial más alta a lo largo de una ruta de reacción. En un estado de transición que tiende a tener una existencia fugaz, las moléculas reactivas en colisión proceden a formar productos. En esta descripción, a veces cuando se describe un sustrato en un proceso, el estado intermedio y de transición también pueden ser adecuados para el proceso. En tales situaciones, el sustrato, el intermedio y el estado de transición pueden denominarse colectivamente "ligandos". En algunos casos, se generan múltiples intermedios en la transformación catalítica de un sustrato. En ciertas realizaciones, se sabe que la especie de ligando (sustrato o estado intermedio o de transición) elegido para el análisis está asociada con un paso limitante de la velocidad en la transformación catalítica. Como ejemplo, un sustrato unido covalentemente a un cofactor enzimático puede modificarse químicamente en una etapa de limitación de velocidad. En tal caso, la especie sustrato-cofactor se usa para modelar la interacción.

[**0044]** Como debe ser claro, el concepto de un ligando es más general que el concepto de un "sustrato". Algunos ligandos se unen con un sitio de unión pero no sufren una transformación catalítica. Los ejemplos incluyen ligandos evaluados en el campo del diseño de fármacos. Dichos ligandos pueden ser moléculas pequeñas elegidas por su capacidad para unirse no covalentemente con una biomolécula diana con fines farmacológicos. En algunos casos, se evalúa la capacidad de un ligando para potenciar, activar o inhibir el comportamiento natural de una biomolécula.

[**0045]** Como se usa en este documento, "biomolécula" y "molécula biológica" se refiere a una molécula que se encuentra generalmente en un organismo biológico. En algunas realizaciones, las moléculas biológicas comprenden macromoléculas biológicas poliméricas que tienen múltiples subunidades (es decir, "biopolímeros"). Las biomoléculas típicas incluyen, entre otras, moléculas que comparten algunas características estructurales con polímeros naturales como los ARN (formados a partir de subunidades de nucleótidos), ADN (formados a partir de subunidades de nucleótidos), y péptidos o polipéptidos (formados a partir de subunidades de aminoácidos), que incluyen, por ejemplo, ARN, análogos de ARN, ADN, análogos de ADN, polipéptidos, análogos de polipéptidos, ácidos nucleicos peptídicos (ANP), combinaciones de ARN y ADN (por ejemplo, quimeraplastos) o similares. No se pretende que las biomoléculas se limiten a ninguna molécula en particular, ya que cualquier molécula biológica adecuada encuentra uso en la presente divulgación, que incluye pero no se limita a, por ejemplo, lípidos, carbohidratos u otras moléculas orgánicas que están hechas por una o más moléculas genéticamente codificables (por ejemplo, una o más enzimas o vías enzimáticas) o similares. De particular interés para algunos aspectos de esta descripción son las biomoléculas que tienen sitios de unión que interactúan con un ligando para efectuar una transformación química o biológica, por ejemplo, catálisis de un sustrato, activación de biomolécula o inactivación de la biomolécula.

[**0046]** En algunas realizaciones, una "propiedad beneficiosa" o "actividad" es un aumento o disminución en uno o más de los siguientes: velocidad catalítica (k_{cat}), la afinidad de unión al sustrato (K_M), la eficiencia catalítica (K_{cat}/K_M), especificidad de sustrato, quimioselectividad, regioselectividad, estereoselectividad, estereoespecificidad, especificidad de ligando, agonismo del receptor, antagonismo del receptor, conversión de un cofactor, estabilidad de oxígeno, nivel de expresión de proteínas, solubilidad, termoactividad, termoestabilidad, actividad de pH, estabilidad de pH (por ejemplo, a pH alcalino o ácido), inhibición de la glucosa y/o resistencia a los inhibidores (p. ej., ácido acético, lectinas, ácidos tánicos y compuestos fenólicos) y proteasas. Otras actividades deseadas pueden incluir un perfil alterado en respuesta a un estímulo particular; por ejemplo, temperatura alterada y/o perfiles de pH. En el contexto del diseño racional de ligandos, la optimización de la inhibición covalente dirigida (TCI) es un tipo de actividad. En

algunas realizaciones, dos o más variantes seleccionadas como se describe en el presente documento actúan sobre el mismo sustrato pero difieren con respecto a una o más de las siguientes actividades: velocidad de formación del producto, porcentaje de conversión de un sustrato a un producto, selectividad y/o porcentaje conversión de un cofactor. No se pretende que la presente divulgación se limite a ninguna propiedad beneficiosa particular y/o actividad deseada.

[**0047]** En algunas realizaciones, "actividad" se utiliza para describir el concepto más limitado de la capacidad de una enzima para catalizar la facturación de un sustrato en un producto. Una característica enzimática relacionada es su "selectividad" para un producto particular tal como un enantiómero o producto regioselectivo. La definición amplia de "actividad" presentada aquí incluye selectividad, aunque convencionalmente la selectividad a veces se ve como distinta de la actividad enzimática.

[**0048]** Los términos "proteína", "polipéptido" y "péptido" se usan de forma intercambiable para denotar un polímero de al menos dos amino ácidos unidos covalentemente mediante un enlace amida, independientemente de la longitud o modificación post-translacional (por ejemplo, glicosilación, fosforilación, lipidación, miristilación, ubiquitinación, etc.). En algunos casos, el polímero tiene al menos aproximadamente 30 residuos de aminoácidos, y generalmente al menos aproximadamente 50 residuos de aminoácidos. Más típicamente, contienen al menos aproximadamente 100 residuos de aminoácidos. No se pretende que la presente invención se limite a secuencias de aminoácidos de cualquier longitud específica. Los términos incluyen composiciones consideradas convencionalmente como fragmentos de proteínas o péptidos de longitud completa. Se incluyen dentro de esta definición los aminoácidos D y L y las mezclas de aminoácidos D y L. Los polipéptidos descritos aquí no están restringidos a los aminoácidos codificados genéticamente. De hecho, además de los aminoácidos codificados genéticamente, los polipéptidos descritos en el presente documento pueden estar formados, en su totalidad o en parte, por aminoácidos no codificados de origen natural y/o sintéticos. En algunas realizaciones, un polipéptido es una porción del polipéptido ancestral o parental de longitud completa, que contiene adiciones o deleciones de aminoácidos (por ejemplo, huecos), y/o sustituciones en comparación con la secuencia de aminoácidos del polipéptido parental de longitud completa, mientras se mantiene la actividad funcional (p. ej., actividad catalítica).

[**0049]** Como se usa en este documento, el término "de tipo silvestre" o "tipo salvaje" (WT) se refiere a organismos, las enzimas y/u otras proteínas de origen natural (por ejemplo, enzimas no recombinantes). Un sustrato o ligando que reacciona con una biomolécula de tipo salvaje a veces se considera un sustrato o ligando "nativo".

[**0050]** Como se usa en este documento, los términos "variante", "mutante", "secuencia mutante" y "variante de secuencia" se refieren a una secuencia biológica que difiere en algún aspecto de una secuencia estándar o de referencia (por ejemplo, en algunas realizaciones, una secuencia parental). La diferencia puede denominarse "mutación". En algunas realizaciones, un mutante es un polipéptido o secuencia de polinucleótidos que ha sido alterada por al menos una sustitución, inserción, cruce, deleción y/u otra operación genética. Para los propósitos de la presente divulgación, los mutantes y variantes no están limitados a un método particular por el cual se generan. En algunas realizaciones, una secuencia mutante o variante tiene actividades o propiedades aumentadas, disminuidas o sustancialmente similares, en comparación con la secuencia parental. En algunas realizaciones, el polipéptido variante comprende uno o más residuos de aminoácidos que han sido mutados, en comparación con la secuencia de aminoácidos del polipéptido de tipo salvaje (por ejemplo, un polipéptido original). En algunas realizaciones, uno o más residuos de aminoácidos del polipéptido se mantienen constantes, son invariables o no están mutados en comparación con un polipéptido original en los polipéptidos variantes que forman una pluralidad de polipéptidos. En algunas realizaciones, el polipéptido original se usa como base para generar variantes con estabilidad, actividad o cualquier otra propiedad deseada mejorada.

[**0051]** Como se usa en este documento, los términos "variante de enzima" y "enzima de la variante" se usan en referencia a las enzimas que son similares a una enzima de referencia, en particular en su función, pero que tienen mutaciones en su secuencia de aminoácidos que las hacen diferentes en secuencia del tipo salvaje u otra enzima de referencia. Las variantes de enzimas pueden hacerse mediante una amplia variedad de técnicas de mutagénesis diferentes bien conocidas por los expertos en la materia. Además, los kits de mutagénesis también están disponibles en muchos proveedores comerciales de biología molecular. Los métodos están disponibles para realizar sustituciones específicas en aminoácidos definidos (dirigidos al sitio), mutaciones específicas o aleatorias en una región localizada del gen (regioespecífica) o mutagénesis aleatoria sobre todo el gen (por ejemplo, mutagénesis de saturación). Los expertos en la técnica conocen numerosos métodos adecuados para generar variantes enzimáticas, que incluyen, pero no se limitan a, mutagénesis dirigida al sitio del ADN o bicatenario ADN usando PCR, mutagénesis en casete, síntesis génica, PCR propensa a errores, barajado y mutagénesis de saturación química, o cualquier otro método adecuado conocido en la técnica. Después de que se producen las variantes, pueden seleccionarse para la propiedad deseada (por ejemplo, alta o aumentada; o actividad baja o reducida, mayor estabilidad térmica y/o alcalina, etc.).

[**0052]** Un "panel de enzimas" es un grupo de enzimas seleccionado de tal manera que cada miembro del panel cataliza la misma reacción química. En algunas realizaciones, los miembros del panel pueden voltear colectivamente múltiples sustratos, cada uno de los cuales experimenta la misma reacción. A menudo, los miembros del panel son elegidos para entregar eficientemente múltiples sustratos. En algunos casos, los paneles están disponibles comercialmente. En otros casos, son propiedad de una entidad. Por ejemplo, un panel puede incluir varias enzimas identificadas como éxitos en un procedimiento de detección. En ciertas realizaciones, uno o más miembros de un

panel existen solo como una representación computacional. En otras palabras, la enzima es una enzima virtual.

[**0053]** Un "modelo" es una representación de la estructura de una biomolécula o ligando. A veces se proporciona como una colección de posiciones tridimensionales para los átomos o restos de la entidad representada. Los modelos a menudo contienen representaciones producidas computacionalmente de los sitios de unión u otros aspectos de las variantes enzimáticas. Los ejemplos de modelos relevantes para las realizaciones en el presente documento se producen a partir de modelos de homología, enhebrado de proteínas o modelos de proteínas *ab initio* usando una rutina tal como Rosetta (rosettacommons.org/software/) o simulaciones de Molecular Dynamics.

[**0054]** Un "modelo de homología" es un modelo tridimensional de una proteína o parte de una proteína que contiene al menos el sitio de unión de un ligando de bajo consideración. El modelo de homología se basa en la observación de que las estructuras de proteínas tienden a conservarse entre las proteínas homólogas. Un modelo de homología proporciona posiciones tridimensionales de residuos que incluyen cadena principal y cadenas laterales. El modelo se genera a partir de una plantilla de estructura de una proteína homóloga que probablemente se parezca a la estructura de la secuencia modelada. En algunas realizaciones, una plantilla de estructura se usa en dos pasos: "alinear secuencia con plantillas" y "construir modelos de homología".

[**0055]** La etapa "alinear secuencia a plantillas" alinea la secuencia modelo para una o más secuencias modelo de estructura y prepara una alineación de la secuencia de entrada para la construcción del modelo de homología. La alineación identifica huecos y otras regiones de disimilitud entre la secuencia del modelo y la(s) secuencia(s) de plantilla de estructura.

[**0056]** La etapa de "construcción de modelos de homología" utiliza características estructurales de la plantilla de estructura para derivar las restricciones espaciales que, a su vez, se utilizan para generar, por ejemplo, estructuras de proteínas modelo usando gradiente conjugado y procedimientos de optimización de recocido simulado. Las características estructurales de la plantilla se pueden obtener de una técnica tal como RMN o cristalografía de rayos x. Se pueden encontrar ejemplos de tales técnicas en el artículo de revisión, "A Guide to Template Based Structure Prediction," by Qu X, Swanson R, Day R, Tsai J. Curr Protein Pept Sci. Junio de 2009; 10 (3): 270-85.

[**0057]** El término "conformación activa" se usa en referencia a una conformación de una proteína (por ejemplo, una enzima) que permite la proteína para hacer que un sustrato se someta a una transformación química (por ejemplo, una reacción catalítica).

[**0058]** Una "postura activo" es una en donde es probable que un ligando padezca una transformación catalítica o realice algún papel deseado tal como la unión de forma covalente con el sitio de unión.

[**0059]** El término "secuencia" se utiliza aquí para referirse a la orden y la identidad de cualquiera de las secuencias biológicas incluyendo, pero no limitado a, un genoma entero, todo el cromosoma, el segmento de cromosoma, colección de secuencias de genes para interactuar genes, gen, secuencia de ácido nucleico, proteínas, péptidos, polipéptidos, polisacáridos, etc. En algunos contextos, una "secuencia" se refiere al orden e identidad de los residuos de aminoácidos en una proteína (es decir, una secuencia de proteínas o cadena de caracteres de proteínas) o al orden e identidad de nucleótidos en un ácido nucleico (es decir, una secuencia de ácido nucleico o cadena de caracteres de ácido nucleico). Una secuencia puede estar representada por una cadena de caracteres. Una "secuencia de ácido nucleico" se refiere al orden y la identidad de los nucleótidos que comprenden un ácido nucleico. Una "secuencia de proteína" se refiere al orden e identidad de los aminoácidos que comprenden una proteína o péptido.

[**0060]** "Codón" se refiere a una secuencia específica de tres nucleótidos consecutivos que es parte del código genético y que especifica un aminoácido particular en una proteína o inicia o detiene la síntesis de proteínas.

[**0061]** El término "gen" se utiliza ampliamente para referirse a cualquier segmento de ADN u otro ácido nucleico asociado con una función biológica. Por lo tanto, los genes incluyen secuencias de codificación y, opcionalmente, las secuencias requeridas para su expresión. Los genes también incluyen opcionalmente segmentos de ácido nucleico no expresados que, por ejemplo, forman secuencias de reconocimiento para otras proteínas. Los genes se pueden obtener de una variedad de fuentes, incluida la clonación de una fuente de interés o sintetizar a partir de información de secuencia conocida o predicha, y pueden incluir secuencias diseñadas para tener los parámetros deseados.

[**0062]** Un "resto" es una parte de una molécula que puede incluir cualquiera de los grupos funcionales enteros o partes de grupos funcionales como subestructuras, mientras que los grupos funcionales son grupos de átomos o enlaces dentro de las moléculas que son responsables de las reacciones químicas características de esas moléculas.

[**0063]** "Cribado" se refiere al proceso en donde se determinan una o más propiedades de una o más biomoléculas. Por ejemplo, los procesos de detección típicos incluyen aquellos en los que se determinan una o más propiedades de uno o más miembros de una o más bibliotecas. El cribado se puede realizar computacionalmente utilizando modelos computacionales de biomoléculas y el entorno virtual de las biomoléculas. En algunas realizaciones, se proporcionan sistemas virtuales de detección de proteínas para enzimas seleccionadas de actividad y selectividad deseadas.

[**0064]** Un "sistema de expresión" es un sistema para expresar una proteína o péptido codificado por un gen u otro ácido nucleico.

[**0065]** "Evolución dirigida", "evolución guiada" o "evolución artificial" se refiere a procesos *in silico, in vitro* o *in vivo* de cambio artificial de una o más secuencias de biomoléculas (o una cadena de caracteres que representa esa secuencia) mediante selección artificial, mutación, recombinación u otra manipulación. En algunas realizaciones, la evolución dirigida ocurre en una población reproductiva en la que (1) hay variedades de individuos, (2) algunas variedades que tienen información genética heredable, y (3) algunas variedades difieren en aptitud. El éxito reproductivo está determinado por el resultado de la selección de una propiedad predeterminada, como una propiedad beneficiosa. La población reproductiva puede ser, por ejemplo, una población física en un proceso *in vitro* o una población virtual en un sistema informático en un proceso *in silico.*

[**0066]** Los métodos de evolución dirigida pueden aplicarse fácilmente a los polinucleótidos para generar bibliotecas de variantes que pueden expresarse, seleccionarse y analizarse. La mutagénesis y los métodos de evolución dirigida son bien conocidos en la técnica (*véanse,* por ejemplo, las patentes de los Estados Unidos N^{os} 5,605,793, 5,830,721, 6,132,970, 6,420,175, 6,277,638, 6,365,408, 6,602,986, 7,288,375, 6,287,861, 6,297,053, 6,576,467, 6,444,468, 6,11,717, 6,179, 6,171, 5,11,717, 6,171, 5,11,717, 6,180,406, 6,291,242, 6,995,017, 6,395,547, 6,506,602, 6,519,065, 6,506,603, 6,413,774, 6,573,098, 6,323,030, 6,344,356, 6,372,497, 7,868,138, 5,834,252, 5,928,905, 6,489,146, 6,096,548, 6,387,702, 6,391,552, 6,358,742, 6,482,647, 6,335,160, 6,653,072, 6,355,484, 6, 03,344, 6,319,713, 6,613,514, 6,455,253, 6,579,678, 6,586,182, 6,406,855, 6,946,296, 7,534,564, 7,776,598, 5,837,458, 6,391,640, 6,309,883, 7,105,297, 7,795,030, 6,326,204, 6,251,674, 6,716,631, 6,528,311, 6,287,862, 6,335,198, 6,352,859, 6,379,964, 7,148,054, 7,629,170 7,620,500, 6,365,377, 6,358,740, 6,406,910, 6,413,745, 6,436,675, 6,961,664, 7,430,477, 7,873,499, 7,702,464, 7,783,428, 7,747,391, 7,747,393, 7,751,986, 6,376,246, 6,426,542, 6,423,642, 6,426,642, 6,426,242 714, 6,521,453, 6,368,861, 7,421,347, 7,058,515, 7,024,312, 7,620,502, 7,853,410, 7,957,912, 7,904,249, y todas las contrapartes no estadounidenses relacionadas; Ling y col., Anal. Biochem, 254 (2): 157-78 [1997]; Dale y col., Meth. Mol. Biol., 57: 369-74 [1996]; Smith, Ann. Rev. Genet., 19: 423-462 [1985]; Botstein y col., Science, 229: 1193-1201 [1985]; Carter, Biochem. J., 237: 1-7 [1986]; Kramer y col., Cell, 38: 879-887 [1984]; Wells y col., Gene, 34: 315-323 [1985]; Minshull y col., Curr. Op. Chem Biol., 3: 284-290 [1999]; Christians et al., Nat. Biotechnol., 17: 259-264 [1999]; Crameri y col., Nature, 391: 288-291 [1998]; Crameri y col., Nat. Biotechnol., 15: 436-438 [1997]; Zhang y col., Proc. Nat. Acad. Sci. Estados Unidos, 94: 4504-4509 [1997]; Crameri y col., Nat. Biotechnol., 14: 315-319 [1996]; Stemmer, Nature, 370: 389-391 [1994]; Stemmer, Proc. Nat. Acad. Sci. Estados Unidos, 91:10747-10751 [1994]; WO 95/22625; WO 97/0078; WO 97/35966; WO 98/27230; WO 00/42651; WO 01/75767; y WO 2009/152336).

[**0067]** En ciertas realizaciones, los métodos de evolución dirigida generan bibliotecas de variante de proteína por recombinación de genes que codifican variantes desarrolladas a partir de una proteína de matriz, así como por recombinación de genes que codifican variantes de una biblioteca padre de variante de la proteína. Los métodos pueden emplear oligonucleótidos que contienen secuencias o subsecuencias que codifican al menos una proteína de una biblioteca de variantes parentales. Algunos de los oligonucleótidos de la biblioteca de variantes parentales pueden estar estrechamente relacionados, difiriendo solo en la elección de codones para aminoácidos alternativos seleccionados para ser variados por recombinación con otras variantes. El método puede realizarse durante uno o varios ciclos hasta que se logren los resultados deseados. Si se utilizan múltiples ciclos, cada uno típicamente implica un paso de selección para identificar aquellas variantes que tienen un rendimiento aceptable o mejorado y son candidatos para su uso en al menos un ciclo de recombinación posterior. En algunas realizaciones, la etapa de selección implica un sistema virtual de detección de proteínas para determinar la actividad catalítica y la selectividad de las enzimas para los sustratos deseados.

[**0068]** En algunas realizaciones, los métodos de evolución dirigida generan variantes de la proteína por mutagénesis dirigida localizadas en los residuos definidos. Estos residuos definidos se identifican típicamente mediante análisis estructural de sitios de unión, análisis de química cuántica, análisis de homología de secuencia, modelos de actividad de secuencia, etc. Algunas realizaciones emplean mutagénesis de saturación, en la que se intenta generar todas las mutaciones posibles (o lo más cercanas posible) en un sitio específico, o región estrecha de un gen.

[**0069]** El "reordenamiento" y el "reordenamiento genético" son tipos de métodos de evolución dirigida que recombinan una colección de fragmentos de los polinucleótidos parentales a través de una serie de ciclos de extensión de cadena. En ciertas realizaciones, uno o más de los ciclos de extensión de cadena es autocebante; es decir, realizado sin la adición de cebadores que no sean los fragmentos mismos. Cada ciclo implica el recocido de fragmentos monocatenarios mediante hibridación, el alargamiento posterior de fragmentos recocidos a través de la extensión de la cadena y la desnaturalización. En el transcurso de la combinación aleatoria, una cadena de ácido nucleico en crecimiento se expone típicamente a múltiples parejas de recocido diferentes en un proceso a veces denominado "cambio de plantilla", que implica cambiar un dominio de ácido nucleico de un ácido nucleico con un segundo dominio de un segundo ácido nucleico (es decir, los ácidos nucleicos primero y segundo sirven como plantillas en el procedimiento de reordenamiento).

[**0070]** El cambio de plantilla frecuentemente produce secuencias quiméricas, que resultan de la introducción de cruces entre fragmentos de diferentes orígenes. Los cruces se crean a través de recombinaciones conmutadas por plantilla

durante los múltiples ciclos de recocido, extensión y desnaturalización. Por lo tanto, la combinación aleatoria conduce típicamente a la producción de secuencias de polinucleótidos variantes. En algunas realizaciones, las secuencias variantes comprenden una "biblioteca" de variantes (es decir, un grupo que comprende múltiples variantes). En algunas realizaciones de estas bibliotecas, las variantes contienen segmentos de secuencia de dos o más de polinucleótidos parentales.

[**0071]** Cuando se emplean dos o más polinucleótidos parentales, los polinucleótidos parentales individuales son suficientemente homólogos que los fragmentos de diferentes padres hibridan en las condiciones de recocido empleados en los ciclos de reordenamiento. En algunas realizaciones, la combinación aleatoria permite la recombinación de polinucleótidos parentales que tienen niveles de homología relativamente limitados/bajos. A menudo, los polinucleótidos parentales individuales tienen dominios distintos y/o únicos y/u otras características de secuencia de interés. Cuando se usan polinucleótidos parentales que tienen características de secuencia distintas, la combinación aleatoria puede producir polinucleótidos variantes muy diversos.

[**0072]** Diversas técnicas de reordenamiento son conocidas en la técnica. Véanse, por ejemplo, las patentes de los Estados Unidos números 6.917.882, 7.776.598, 8.029.988, 7.024.312 y 7.795.030.

[**0073]** Algunas técnicas de evolución dirigida emplean "empalme de genes por extensión de solapamiento" o "SOEing de gen", que es un método basado en PCR de recombinación de secuencias de ADN sin depender de sitios de restricción y de generar directamente fragmentos de ADN mutados *in vitro.* En algunas implementaciones de la técnica, las PCR iniciales generan segmentos de genes superpuestos que se utilizan como plantilla de ADN para una segunda PCR para crear un producto de longitud completa. Los cebadores de PCR internos generan extremos 3' complementarios superpuestos en segmentos intermedios e introducen sustituciones, inserciones o deleciones de nucleótidos para el empalme de genes. Las cadenas superpuestas de estos segmentos intermedios se hibridan en la región 3' en la segunda PCR y se extienden para generar el producto de longitud completa. En diversas aplicaciones, el producto de longitud completa se amplifica mediante cebadores flanqueantes que pueden incluir sitios de enzimas de restricción para insertar el producto en un vector de expresión con fines de clonación. Ver, por ejemplo, Horton, et al., Biotechniques, 8 (5): 528-35 [1990]. La "mutagénesis" es el proceso de introducir al menos una mutación en una secuencia estándar o de referencia tal como un ácido nucleico o polipéptido original.

[**0074]** La mutagénesis dirigida al sitio es un ejemplo de una técnica útil para introducir mutaciones, aunque cualquier método adecuado encuentra uso. Por lo tanto, de forma alternativa o adicional, los mutantes pueden proporcionarse mediante síntesis génica, mutagénesis aleatoria saturada, bibliotecas combinatorias semisintéticas de residuos, recombinación de secuencia recursiva ("RSR") (véase, por ejemplo, la Solicitud de Patente de EE.UU. Nº 2006/0223143), mezcla aleatoria de genes, PCR propensa a errores y/o cualquier otro método adecuado.

[**0075]** Un ejemplo de un procedimiento de mutagénesis de saturación adecuado se describe en la Solicitud de Patente de EE.UU. Publ. Nº 2010/0093560.

[**0076]** Un "fragmento" es cualquier porción de una secuencia de nucleótidos o aminoácidos. Los fragmentos pueden producirse usando cualquier método adecuado conocido en la técnica, que incluye pero no se limita a escindir un polipéptido o secuencia de polinucleótidos. En algunas realizaciones, los fragmentos se producen usando nucleasas que escinden polinucleótidos. En algunas realizaciones adicionales, los fragmentos se generan usando técnicas de síntesis química y/o biológica. En algunas realizaciones, los fragmentos comprenden subsecuencias de al menos una secuencia parental, generada usando alargamiento de cadena parcial de ácido(s) nucleico(s) complementario(s). En algunas realizaciones que implican técnicas de *silico,* los fragmentos virtuales se generan computacionalmente para imitar los resultados de fragmentos generados por técnicas químicas y/o biológicas. En algunas realizaciones, los fragmentos de polipéptidos exhiben la actividad del polipéptido de longitud completa, mientras que en algunas otras realizaciones, los fragmentos de polipéptidos no tienen la actividad exhibida por el polipéptido de longitud completa.

[**0077]** "Polipéptido parental", "polinucleótido parental", "ácido nucleico parental" y "parental" se usan generalmente para referirse al polipéptido de tipo salvaje, polinucleótido de tipo salvaje o una variante usada como punto de partida en un procedimiento de generación de diversidad, como una evolución dirigida. En algunas realizaciones, el padre mismo se produce mediante reordenamiento u otro(s) procedimiento(s) de generación de diversidad. En algunas realizaciones, los mutantes utilizados en la evolución dirigida están directamente relacionados con un polipéptido original. En algunas realizaciones, el polipéptido original es estable cuando se expone a condiciones extremas de temperatura, pH y/o disolvente y puede servir como base para generar variantes para reordenamiento. En algunas realizaciones, el polipéptido parental no es estable a condiciones extremas de temperatura, pH y/o disolvente, y el polipéptido parental se desarrolla para hacer variantes robustas.

[**0078]** Un "ácido nucleico original" codifica un polipéptido parental.

[**0079]** Una "biblioteca" o "población" se refiere a una colección de al menos dos moléculas diferentes, cadenas de caracteres, y/o modelos, tales como secuencias de ácido nucleico (por ejemplo, genes, oligonucleótidos, etc.) o productos de expresión (por ejemplo,, enzimas u otras proteínas) a partir de ellas. Una biblioteca o población generalmente incluye varias moléculas diferentes. Por ejemplo, una biblioteca o población típicamente incluye al

menos aproximadamente 10 moléculas diferentes. Las bibliotecas grandes típicamente incluyen al menos aproximadamente 100 moléculas diferentes, más típicamente al menos aproximadamente 1000 moléculas diferentes. Para algunas aplicaciones, la biblioteca incluye al menos aproximadamente 10000 o más moléculas diferentes. Sin embargo, no se pretende que la presente invención se limite a un número específico de moléculas diferentes. En ciertas realizaciones, la biblioteca contiene una variante numérica o ácidos nucleicos o proteínas quiméricos producidos por un procedimiento de evolución dirigida.

[**0080]** Dos ácidos nucleicos se "recombinan" cuando las secuencias de cada uno de los dos ácidos nucleicos se combinan para producir ácido(s) nucleico(s) de la progenie. Dos secuencias se recombinan "directamente" cuando ambos ácidos nucleicos son sustratos para recombinación.

[**0081]** El término "selección" se refiere al proceso en donde se identifican una o más biomoléculas como que tiene una o más propiedades de interés. Por lo tanto, por ejemplo, se 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 esto no es necesario. Además, la selección y el cribado pueden ser, y a menudo son, simultáneos. Algunas realizaciones descritas en el presente documento proporcionan sistemas y métodos para seleccionar enzimas de actividad y/o selectividad deseables.

[**0082]** La "secuenciación de próxima generación" y la "secuenciación de alto rendimiento" son técnicas de secuenciación que paralelizan el proceso de secuenciación, produciendo miles o millones de secuencias a la vez. Los ejemplos de métodos adecuados de secuenciación de próxima generación incluyen, entre otros, secuenciación en tiempo real de una sola molécula (p. ej., Pacific Biosciences, Menlo Park, California), secuenciación de semiconductores iónicos (p. ej., Ion Torrent, South San Francisco, California), secuenciación de pirosis (p. ej., 454, Branford, Connecticut), secuenciación por ligadura (p. ej., secuenciación SOLiD de Life Technologies, Carlsbad, California), secuenciación por síntesis y reversible terminador (p. ej., Illumina, San Diego, California), tecnologías de imágenes de ácido nucleico tales como microscopía electrónica de transmisión y similares.

[**0083]** Una "variable dependiente" ("DV") representa una salida o efecto, o se prueba para ver si es el efecto. Las "variables independientes" ("IVs") representan las entradas o causas, o se prueban para ver si son la causa. Se puede estudiar una variable dependiente para ver si varía y cuánto varía a medida que varían las variables independientes.

[**0084]** En el modelo lineal estocástico sencillo

$$y_i = a + bx_i + e_i$$

donde el término y_i es el $i^{\text{ésimo}}$ valor de la variable dependiente y x_i es $i^{\text{ésimo}}$ valor de la variable independiente (IV). El término e_i se conoce como "error" y contiene la variabilidad de la variable dependiente no explicada por la variable independiente.

[**0085]** Una variable independiente (IV) también se conoce como una "variable de predictor", "regresor", "variable controlada", "variable manipulada", "variable explicativa", o "variable de entrada".

[**0086]** El término "coeficiente" se refiere a un valor escalar multiplicado por una variable dependiente o una expresión que contiene una variable dependiente.

[**0087]** Los términos "ortogonal" y "ortogonalidad" se refieren a una variable independiente que no está correlacionada con otras variables independientes en un modelo u otra relación.

[**0088]** El término "modelo de actividad de la secuencia" se refiere a cualquiera de los modelos matemáticos que describen la relación entre las actividades, características o propiedades de las moléculas biológicas, por un lado, y varias secuencias biológicas en la otra mano.

[**0089]** El término "cadena de caracteres" se refiere a una representación de una molécula biológica que conserva la información de secuencia/estructural con respecto a esa molécula. En algunas realizaciones, la cadena de caracteres contiene información sobre mutaciones de secuencia en una biblioteca de variantes. Las cadenas de caracteres de las biomoléculas y la información de actividad para las biomoléculas pueden usarse como un conjunto de entrenamiento para un modelo de actividad de secuencia. Las propiedades no secuenciales de las biomoléculas pueden almacenarse o asociarse de otro modo con cadenas de caracteres para las biomoléculas.

[**0090]** Una "secuencia de referencia" es una secuencia de la que se efectúa la variación de secuencia. En algunos casos, se utiliza una " secuencia de referencia " para definir las variaciones. Tal secuencia puede ser predicha por un modelo para tener el valor más alto (o uno de los valores más altos) de la actividad deseada. En otro caso, la secuencia de referencia puede ser la de un miembro de una biblioteca de variantes de proteínas originales. En ciertas realizaciones, una secuencia de referencia es la secuencia de una proteína parental o ácido nucleico.

[**0091]** La frase "conjunto de entrenamiento" se refiere a un conjunto de datos de la secuencia de actividad u

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

[**0092]** El término "observación" es información acerca de la proteína u otra entidad biológica que puede ser utilizada en un conjunto de entrenamiento para la generación de un modelo como un modelo de actividad de la secuencia. El término "observación" puede referirse a cualquier molécula biológica secuenciada y/o analizada, incluidas las variantes de proteínas. En ciertas realizaciones, cada observación es un valor de actividad y una secuencia asociada para una variante en una biblioteca. Generalmente, cuantas más observaciones se empleen para crear un modelo de actividad de secuencia, mejor será el poder predictivo de ese modelo de actividad de secuencia.

[**0093]** La frase "poder predictivo" se refiere a la capacidad de un modelo para predecir correctamente los valores de una variable dependiente para los datos en diversas condiciones. Por ejemplo, el poder predictivo de un modelo de actividad de secuencia se refiere a la capacidad del modelo para predecir actividad a partir de información de secuencia.

[**0094]** La frase "validación cruzada" se refiere a un método para probar la generalización de la capacidad de un modelo para predecir el valor de la variable dependiente. El método prepara un modelo utilizando un conjunto de datos y prueba el error del modelo utilizando un conjunto diferente de datos. El primer conjunto de datos se ve como un conjunto de entrenamiento, y el segundo conjunto de datos es un conjunto de validación.

[**0095]** La frase "variación sistemática" se refiere a diferentes descriptores de un artículo o conjunto de artículos que se cambian en diferentes combinaciones.

[**0096]** La frase "datos variaron sistemáticamente" se refiere a los datos producidos, derivados, o resultantes de los diferentes descriptores de un artículo o conjunto de artículos que son cambiados en diferentes combinaciones. Se pueden cambiar muchos descriptores diferentes al mismo tiempo, pero en diferentes combinaciones. Por ejemplo, los datos de actividad recopilados de polipéptidos en los que se han cambiado combinaciones de aminoácidos son datos variados sistemáticamente.

[**0097]** La frase "secuencias sistemáticamente variadas" se refiere a un conjunto de secuencias en las que cada residuo se ve en contextos múltiples. En principio, el nivel de variación sistemática se puede cuantificar por el grado en que las secuencias son ortogonales entre sí (es decir, máximamente diferentes en comparación con la media).

[**0098]** El término "alternar" se refiere a la introducción de múltiples tipos de residuos de aminoácidos en una posición específica en las secuencias de proteínas variantes en la biblioteca optimizada.

[**0099]** Los términos "regresión" y "análisis de regresión" se refieren a las técnicas utilizadas para entender cuál de las variables independientes están relacionadas con la variable dependiente, y para explorar las formas de estas relaciones. En circunstancias restringidas, el análisis de regresión puede usarse para inferir relaciones causales entre las variables independientes y dependientes. Es una técnica estadística para estimar las relaciones entre variables. Incluye muchas técnicas para modelar y analizar varias variables, cuando el foco está en la relación entre una variable dependiente y una o más variables independientes. Más específicamente, el análisis de regresión ayuda a comprender cómo cambia el valor típico de la variable dependiente cuando se modifica cualquiera de las variables independientes, mientras que las otras variables independientes se mantienen fijas. Las técnicas de regresión pueden usarse para generar modelos de actividad de secuencia a partir de conjuntos de entrenamiento que comprenden múltiples observaciones, que pueden contener información de secuencia y actividad.

[**0100]** "Mínimos cuadrados parciales" ("PLS") es una familia de métodos que encuentra un modelo de regresión lineal proyectando variables predichas (p. ej., actividades) y las variables observables (p. ej., secuencias) a un nuevo espacio. PLS también se conoce como "proyección a estructuras latentes". Tanto los datos *X* (variables independientes) como *Y* (variables dependientes) se proyectan a nuevos espacios. PLS se utiliza para encontrar las relaciones fundamentales entre dos matrices (*X* e Y). Se usa un modelo de variable latente para modelar las estructuras de covarianza en los espacios *X* e *Y.* Un modelo PLS intentará encontrar la dirección multidimensional en el espacio *X* que explica la dirección máxima de la varianza multidimensional en el espacio *Y.* La regresión PLS es particularmente útil cuando la matriz de predictores tiene más variables que observaciones, y cuando hay una multicolinealidad entre los valores X.

[**0101]** Las variables latentes (a diferencia de las variables observables) son variables que no se observan directamente pero se infieren de variables observadas o medidas directamente. Los modelos matemáticos que tienen como objetivo explicar las variables observadas en términos de variables latentes se denominan modelos de variables latentes.

[**0102]** Un "descriptor" se refiere a algo que sirve para describir o identificar un elemento. Por ejemplo, los caracteres en una cadena de caracteres pueden ser descriptores de aminoácidos en un polipéptido representado por la cadena de caracteres.

[**0103]** En un modelo de regresión, la variable dependiente está relacionada con variables independientes mediante una suma de términos. Cada término incluye un producto de una variable independiente y un coeficiente de regresión asociado. En el caso de un modelo de regresión puramente lineal, los coeficientes de regresión están dados por β en la siguiente forma de expresión:

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i$$

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

[**0104]** La frase "regresión del componente principal" ("PCR") se refiere a un análisis de regresión que utiliza el análisis del componente principal al estimar los coeficientes de regresión. En lugar de retroceder la variable dependiente en las variables independientes directamente, se utilizan los componentes principales de las variables independientes. La PCR generalmente solo usa un subconjunto de los componentes principales en el análisis de regresión.

[**0105]** La frase "análisis de componentes principales" ("PCA") se refiere a un procedimiento matemático que usa una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas llamadas "componentes principales". El número de componentes principales es menor o igual que el número de variables originales. Esta transformación se define de tal manera que el primer componente principal tiene la mayor varianza posible (es decir, representa la mayor variabilidad posible en los datos), y cada componente subsiguiente a su vez tiene la mayor varianza posible bajo la restricción que sea ortogonal a (es decir, no correlacionada con) los componentes anteriores.

[**0106]** Una "red neuronal" es un modelo que contiene un grupo interconectado de elementos de procesamiento o "neuronas" que procesan información usando un enfoque conexionista para la computación. Las redes neuronales se utilizan para modelar relaciones complejas entre entradas y salidas y/o para encontrar patrones en los datos. La mayoría de las redes neuronales procesan datos de manera no lineal, distribuida y paralela. En la mayoría de los casos, las redes neuronales son sistemas adaptativos que cambian su estructura durante una fase de aprendizaje. Las funciones se realizan colectivamente y en paralelo por los elementos de procesamiento, en lugar de utilizar una delimitación clara de las subtareas a las que se asignan varias unidades.

[**0107]** En general, una red neuronal implica una red de elementos de procesamiento simples que exhiben un comportamiento global complejo determinado por las conexiones entre los elementos de procesamiento y los parámetros del elemento. Las redes neuronales se usan con algoritmos diseñados para alterar la fuerza de las conexiones en la red para producir un flujo de señal deseado. La fuerza se altera durante el entrenamiento o el aprendizaje.

[**0108]** Un "algoritmo genético" ("GA") es un proceso que imita procesos evolutivos. Los algoritmos genéticos (AG) son utilizados en una amplia variedad de campos para resolver problemas que no están completamente caracterizados o son demasiado complejos para permitir una caracterización completa, pero para los cuales alguna evaluación analítica está disponible. Es decir, los GA se usan para resolver problemas que pueden evaluarse mediante alguna medida cuantificable para el valor relativo de una solución (o al menos el valor relativo de una solución potencial en comparación con otra). En el contexto de la presente divulgación, un algoritmo genético es un proceso para seleccionar o manipular cadenas de caracteres en una computadora, típicamente donde la cadena de caracteres corresponde a una o más moléculas biológicas (por ejemplo, ácidos nucleicos, proteínas o similares) o datos se usa para entrenar un modelo como un modelo de actividad de secuencia o una máquina de vectores de soporte.

[**0109]** En un ejemplo, un algoritmo genético proporciona y evalúa una población de modelos en una primera generación del algoritmo. Cada modelo incluye múltiples parámetros que describen la relación entre al menos una variable independiente (IV) y una variable dependiente (DV). Una "función de aptitud" evalúa los modelos de miembros de la población y los clasifica en función de uno o más criterios, como la alta actividad deseada o el bajo error de predicción del modelo. Los modelos miembros de la población también se denominan a veces individuos o cromosomas en el contexto de algoritmos genéticos. En algunas realizaciones, la aptitud del modelo se evalúa utilizando el Criterio de Información de Akaike (AIC) o el Criterio de Información Bayesiano (BIC), en donde los individuos que tienen los valores de AIC o BIC más pequeños se eligen como los individuos más aptos. Los modelos de alto rango se seleccionan para su promoción a una segunda generación y/o apareamiento para producir una población de "hijos modelos" para una segunda generación del algoritmo. La población en la segunda generación es evaluada de manera similar por la función de aptitud física, y los miembros de alto rango son promovidos y/o apareados como en la primera generación. El algoritmo genético continúa de esta manera para las generaciones posteriores hasta que se cumpla un "criterio de convergencia", momento en el cual el algoritmo concluye con uno o más individuos de alto rango (modelos).

[**0110]** En otro ejemplo, los "individuos" son secuencias de péptidos variantes y la función de aptitud es la actividad predicha de estos individuos. Cada generación contiene una población de secuencias de péptidos individuales, y se evalúa su aptitud. Los más aptos en una generación son seleccionados para promoción y/o apareamiento para producir una población de próxima generación. Después de varias generaciones, el algoritmo genético puede converger en una población de secuencias peptídicas de alto rendimiento.

[**0111]** Como en el ejemplo anterior, un algoritmo genético a menudo se ejecuta a través de múltiples iteraciones para buscar parámetros óptimos en un espacio de parámetros. Cada iteración del algoritmo genético también se conoce como una "generación" del algoritmo genético. Los modelos en una generación del algoritmo genético forman una "población" para la generación. En el contexto de algoritmos genéticos, los términos "cromosoma" e "individuo" a veces se usan como alias para un modelo o un conjunto de parámetros del modelo en una población. Se usa así porque un modelo de una generación parental pasa sus parámetros (o "genes") a los modelos de una generación infantil, que se asemeja al proceso biológico en donde un cromosoma original pasa sus genes a un cromosoma infantil.

[**0112]** El término "operación genética" ("GO") se refiere a operaciones genéticas biológicas y/o computacionales, en donde todos los cambios en cualquier población de cualquier tipo de cadenas de caracteres (y por lo tanto en cualquier propiedad física de los objetos físicos codificados por tales cadenas) puede describirse como resultado de la aplicación aleatoria y/o predeterminada de un conjunto finito de funciones algebraicas lógicas. Los ejemplos de GO incluyen, entre otros, multiplicación, cruce, recombinación, mutación, ligadura, fragmentación, etc.

[**0113]** El "Criterio de Información de Akaike" ("AIC") es una medida de la bondad relativa de ajuste de un modelo estadístico, y a menudo se usa como criterio para la selección de modelos entre un conjunto finito de modelos. El AIC se basa en el concepto de entropía de la información, que ofrece una medida relativa de la información perdida cuando se utiliza un modelo dado para describir la realidad. Se puede decir que describe la compensación entre el sesgo y la varianza en la construcción del modelo, o hablando en términos generales entre la precisión y la complejidad del modelo. El AIC se puede calcular como:

$$AIC = -2\log_e L + 2k,$$

en donde L es la probabilidad máxima de la función y k es el número de parámetros libres del modelo a estimar.

[**0114]** El "Criterio de Información Bayesiano" ("BIC") es un criterio para la selección de modelos entre un conjunto finito de modelos, y está estrechamente relacionado con AIC. El BIC se puede calcular como: $BIC = -2\log_e L + k\log_e(n)$, en donde n es el número de observaciones de datos. A medida que aumentaba el número de observaciones, BIC a menudo penaliza un número adicional de parámetros libres más que el AIC.

[**0115]** Una "función de verosimilitud" o "probabilidad" de un modelo es una función de los parámetros de un modelo estadístico. La probabilidad de un conjunto de valores de parámetros dados algunos resultados observados es igual a la probabilidad de esos resultados observados dados esos valores de parámetros, es decir, $L(\theta|x) = P(x|\theta)$.

[**0116]** Un "modelo de conjunto" es un modelo cuyos términos incluyen todos los términos de un grupo de modelos, en donde los coeficientes de los términos del modelo de conjunto se basan en los coeficientes ponderados de los términos correspondientes de los modelos individuales del grupo. La ponderación de los coeficientes se basa en el poder predictivo y/o la aptitud de los modelos individuales.

[**0117]** Las "simulaciones de Monte Carlo" son simulaciones que se basan en una gran cantidad de muestreos aleatorios para obtener resultados numéricos que simulan un fenómeno real. Por ejemplo, extrayendo una gran cantidad de variables pseudoaleatorias uniformes del intervalo (0,1] y asignando valores menores o iguales a 0,50 como cabezas y más de 0,50 como colas, es una simulación de Monte Carlo del comportamiento de lanzar una moneda repetidamente.

II. DESCRIPCIONES GENERALES DEL FLUJO DE TRABAJO

A. Flujo de trabajo para una ronda de evolución dirigida

[0118] En ciertas realizaciones, el flujo de trabajo general hace uso de técnicas tanto *in vitro* como computacionales para controlar un proceso de evolución dirigida. El lado computacional del proceso emplea modelos estructurales y secuencia de modelos de actividad.

[**0119]** Cada ronda de evolución dirigida emplea un nuevo conjunto de modelos estructurales y un nuevo modelo de actividad de la secuencia. Además, En cada ronda, las variantes de biomoléculas identificadas para su posterior análisis se evalúan utilizando modelos estructurales tridimensionales de las variantes. La información de los modelos estructurales se combina con las secuencias de las variantes y los datos de ensayo (actividad) para generar un gran conjunto de datos sin filtrar. Por lo general, una parte del conjunto de datos se utiliza como conjunto de entrenamiento.

Para la ronda actual de evolución dirigida, el conjunto de entrenamiento entrena un modelo de actividad de secuencia, que luego identifica las variantes de biomoléculas para la próxima ronda de evolución dirigida.

[**0120]** En ciertas realizaciones, se emplean uno o más algoritmos genéticos (GA) para evaluar los datos combinados no filtrados proporcionados al comienzo de cada ronda de evolución dirigida. Los GA identifican un subconjunto de la información contenida en el conjunto de datos sin filtrar, que se utiliza como variables independientes para entrenar un nuevo modelo de actividad de secuencia. La actividad es la variable dependiente; el modelo de actividad de secuencia proporciona actividad en función de variables independientes identificadas durante el filtrado. En diversas realizaciones, el modelo de actividad de secuencia es un modelo no lineal. En ciertas realizaciones, el modelo de actividad de secuencia es un hiperplano en un espacio n-dimensional, que puede ser generado por una máquina de vectores de soporte.

[**0121]** En un ejemplo representado en la Figura 1A, un flujo de trabajo de evolución dirigida se desarrolla como sigue. Inicialmente, la información se recopila para múltiples variantes de biomoléculas. Cada una de estas variantes puede haber sido identificada en una ronda previa de evolución dirigida. Si el proyecto acaba de comenzar (es decir, no hay rondas previas de evolución dirigida), las variantes se obtienen de una fuente diferente, como un panel de biomoléculas que se sabe que tienen propiedades potencialmente interesantes. A veces, las variantes de la primera ronda se eligen para abarcar un rango relativamente amplio de secuencia y/o espacio de actividad.

[**0122]** Después de que se han identificado las variantes, un sistema de evaluación obtiene varios tipos de información para cada variante. En particular, se determina al menos una actividad de interés y la secuencia de cada variante. En algunas realizaciones, la secuencia se representa como una colección de mutaciones de la secuencia de tipo salvaje u otra secuencia de referencia. En algunas realizaciones, la actividad se almacena como un valor numérico que tiene unidades definidas. En algunas realizaciones, los valores de actividad están normalizados. Si no se conoce la secuencia de una variante dada, se puede obtener secuenciando una muestra física de la variante.

[**0123]** Además de los datos de actividad de secuencia, se genera un modelo estructural para cada biomolécula de variante. En ciertas realizaciones, los modelos estructurales son modelos de homología. Los modelos estructurales se evalúan computacionalmente para obtener datos adicionales que se combinan con los datos de secuencia y actividad para cada variante. En algunas implementaciones, el modelo estructural de cada variante se usa para identificar una energía de interacción de un ligando con el sitio del receptor de la biomolécula y/o uno o más parámetros que describen la geometría del ligando en el sitio del receptor. Dicha geometría puede incluir distancias entre átomos del ligando y átomos de un resto residual en el sitio de unión y/o átomos de un resto cofactor en el sitio de unión. Ciertos ejemplos se presentan a continuación.

[**0124]** El conjunto de datos sin filtrar incluye datos de secuencia y actividad para cada variante y típicamente incluye una variedad de datos adicionales para cada variante. Como se describe en este documento, estos datos adicionales se derivan de los modelos estructurales para cada variante. Además, estos datos adicionales incluyen típicamente (i) interacción o energías de unión entre el ligando en consideración y el sitio de unión de cada variante y/o (ii) descriptores estructurales/geométricos que caracterizan la interacción del ligando con el receptor. Ver el bloque 103 de la Figura 1A.

[**0125]** Se ha descubierto que el conjunto de datos sin filtrar sin procesar no siempre es óptimo para entrenar un nuevo modelo de actividad de secuencia. Por el contrario, un subconjunto filtrado del conjunto de datos sin procesar combinados generalmente proporciona un modelo de actividad de secuencia más útil. Por lo tanto, el conjunto de datos sin procesar del bloque 103 se filtra como se ilustra en los bloques 105 y 107.

[**0126]** El filtrado se puede realizar mediante cualquier técnica adecuada. Como se describe más detalladamente a continuación, una técnica opcional elimina ciertos tipos de parámetros obtenidos de los modelos estructurales de las variantes (por ejemplo, ciertas distancias de átomo de sustrato a átomo de residuo). Bloque 105. Como ejemplo, el conjunto de datos sin filtrar puede contener diez características geométricas disponibles del ligando en el sitio de unión del receptor, pero el filtrado elimina tres de estos, de modo que se utiliza un subconjunto de solo siete parámetros en el conjunto de entrenamiento. Estos parámetros, junto con la secuencia, sirven como variables independientes en un modelo de actividad de secuencia entrenado en el conjunto de entrenamiento. Alternativa o adicionalmente, el filtrado puede eliminar variantes que tienen valores de una o más de las variables independientes que están fuera de un rango o por debajo de un umbral determinado como útil para producir el modelo de actividad de secuencia. Bloque 107. En ciertas realizaciones, las variables independientes filtradas de esta manera se derivan del modelo estructural.

[0127] Como se ilustra en un bloque 109, después de que el conjunto de datos sin procesar se filtró correctamente, se usa para generar el modelo de actividad de secuencia. Como se mencionó, el modelo de actividad de secuencia puede ser un modelo no lineal tal como un hiperplano en un espacio tridimensional determinado por una máquina de vectores de soporte. Después de que se genera el modelo de actividad de secuencia, se utiliza para ayudar a identificar variantes de alto rendimiento para una próxima ronda de evolución dirigida. Véase el bloque 111. En una realización, el modelo de actividad de secuencia entrenada se usa con un algoritmo genético (GA) para seleccionar una pluralidad de secuencias variantes que probablemente tengan propiedades beneficiosas. Las variantes seleccionadas se utilizan en la siguiente ronda de evolución dirigida. En dicha próxima ronda, las variantes seleccionadas con el modelo de

actividad de secuencia se tratan como se describió anteriormente (bloques 103, opcionalmente 105, 107 y 109). Sin embargo, primero se analizan para producir un nuevo conjunto de datos sin procesar. Véase el bloque 113. En ciertas realizaciones, las variantes se producen físicamente y se analizan para determinar su actividad. Esto proporciona algunos de los datos en bruto. Las variantes también se modelan estructuralmente para determinar los valores de energía de interacción y los valores de geometría de unión de ligando para cada uno de los tipos de energía y tipos de geometría utilizados en una ronda anterior de evolución dirigida. Se puede emplear un acoplador para generar valores para estos tipos de datos. Si es necesario, una o más de las variantes se secuencian para completar los datos sin procesar.

[**0128]** Las rondas de evolución dirigida continúan de esta manera hasta que una o más rondas muestran una mejora limitada o cumplen con otros criterios de convergencia. Se concluye el proyecto de evolución dirigida. En la Figura 1A, la verificación de criterios de convergencia se ilustra mediante un bloque de decisión 115.

B. Flujo de trabajo de generación de modelo

[0129] Como se indicó anteriormente, algunas implementaciones filtran un conjunto de datos sin procesar antes de entrenar un modelo de actividad de secuencia. El filtrado puede eliminar ciertos tipos de variables de los datos sin procesar. Cada tipo de variable es una variable independiente potencial para el modelo de actividad de secuencia. Alternativamente, o además, el filtrado puede eliminar ciertas variantes que tienen valores de parámetros fuera de los rangos definidos. Se ha encontrado que tal filtrado reduce el ruido producido por modelos entrenados usando los datos. En algunas implementaciones, el filtrado se lleva a cabo utilizando uno o más GA. En ciertas realizaciones, los tipos de datos filtrados a partir de datos brutos están limitados a la energía de interacción entre el ligando y una biomolécula y/o características geométricas del ligando en el sitio de unión de la biomolécula.

[**0130]** La Figura 1B presenta un enfoque para filtrar datos sin procesar. En la realización representada, los datos de tres fuentes se combinan para formar un conjunto de datos brutos 153. Cada variante aporta sus propios datos de las tres fuentes. Los datos combinados incluyen datos de actividad para una interacción ligando-variante. Los datos de actividad, que están representados por un bloque 141, pueden generarse utilizando herramientas de análisis estándar como la cromatografía líquida, la cromatografía de gases, etc. Además, se proporcionan datos de secuencia para las variantes individuales que tienen los datos de actividad deseados (bloque 141). Los datos de secuencia, que están representados por el bloque 143, pueden conocerse con anticipación o pueden determinarse secuenciando los aminoácidos de las variantes o codificando ácidos nucleicos. La secuenciación puede realizarse utilizando cualquiera de las muchas tecnologías de secuenciación disponibles. La secuenciación masiva paralela se usa en algunas realizaciones. Finalmente, los datos estructurales pueden generarse a partir de modelos estructurales de las variantes. Dicha información se puede obtener utilizando no solo los modelos estructurales sino también un programa de acoplamiento (acoplador) que evalúa las posturas de ligando en el sitio de unión del modelo estructural de una variante en consideración. Los datos estructurales brutos contienen datos para muchos tipos de parámetros, incluidos tipos particulares de energía de interacción y distancias de átomo a átomo entre ligandos y cofactores y/o residuos del sitio de unión. Los datos estructurales brutos están representados por el bloque 145 en la Figura 1B.

[**0131]** Las tres fuentes de datos se combinan como se muestra en la Figura 1B para proporcionar los datos sin procesar combinados 153. En ciertas realizaciones, los datos sin procesar combinados se proporcionan en forma de un archivo legible por computadora o un grupo de archivos que están disponibles para más procesamiento mediante una herramienta de filtrado o algoritmo implementado por computadora.

[**0132]** En la realización representada, se muestran dos etapas separadas de filtrado: la función de selección en la etapa 155 y selección de la distancia en la etapa 157. En la realización representada, cada una de estas operaciones de filtrado se realiza utilizando su propio algoritmo genético empleando su propia actividad de la secuencia modelo como función diana. En una realización específica, los modelos de actividad de secuencia se generan usando máquinas de vectores de soporte 159 y 161, como se representa en la Figura 1B. El filtro de selección de características identifica tipos particulares de energía de interacción y/o distancias de átomo a átomo para eliminar del conjunto de datos sin procesar combinados. En esta realización, el concepto de "distancia" incluye otros parámetros geométricos tales como características de posición angular, torsional y general de los átomos de ligando con respecto a los átomos de biomoléculas y/o cofactores. Los tipos de datos identificados se eliminan para todas las variantes que contribuyen al conjunto de datos. Cuando se utiliza un algoritmo genético, el proceso de eliminación puede ser fluido. En otras palabras, uno o más de los tipos de datos eliminados pueden eliminarse solo temporalmente, durante una o más generaciones, durante la ejecución del algoritmo genético de selección de características. A continuación se describen ejemplos de técnicas adecuadas para lograr esto. El filtro de selección de distancia elimina datos de ciertas variantes que contribuyen a los datos sin procesar. Este filtro selecciona ciertos valores de energía y/o distancia que están fuera de los rangos numéricos designados. Cualquier variante que tenga valores de energía y/o distancia fuera de estos rangos tiene sus datos completamente eliminados del conjunto de datos sin procesar. Cuando el filtrado se implementa utilizando un algoritmo genético, los datos variantes eliminados en un punto del proceso pueden reintroducirse, si corresponde, durante la ejecución posterior del algoritmo genético. Por ejemplo, los datos variantes eliminados durante una generación del algoritmo genético pueden reintroducirse en una generación posterior. El proceso se describirá con más detalle a continuación.

[0133] Después de concluir el filtrado como se describe con respecto a los bloques 155 y 157, se entrena un modelo de actividad de secuencia usando los datos filtrados. En algunas implementaciones, la capacitación se lleva a cabo utilizando una máquina de vectores de soporte. El modelo de actividad de secuencia resultante se representa como el bloque 165. Se utiliza como una función diana en un algoritmo genético diferente, que considera y clasifica las secuencias variantes según los valores de actividad pronosticados. El algoritmo genético en cuestión se representa en el bloque 167 de la Figura 1B.

[**0134]** En algunas otras realizaciones, la etapa 155 de selección de características no se realiza. Por lo tanto, no se filtra ninguna característica. En otras palabras, todas las características disponibles se usan para entrenar el modelo de actividad de secuencia 165 para el algoritmo genético de predicción 167. El filtrado solo elimina las variantes que tienen valores de energía o geometría fuera de los rangos identificados. En algunas otras realizaciones, la etapa de selección de características 155, y la etapa de selección de distancia 157, se combinan en una sola etapa de selección, que puede implementarse usando un algoritmo genético. En estas realizaciones, los tipos de características y los valores de características varían en los datos del conjunto de entrenamiento evaluados usando un algoritmo genético.

[**0135]** La Figura 1C presenta un proceso en donde la etapa de selección de características no se realiza o se combina con la etapa de selección de distancia 157. Como se muestra, los datos en bruto 171 se filtran usando un único algoritmo genético 173, que selecciona variantes que tienen uno o más parámetros de geometría restringidos dentro de los rangos elegidos. En un ejemplo, los parámetros de geometría son distancias entre los átomos de un sustrato y los átomos de un residuo o cofactor en un sitio de unión. Por ejemplo, un parámetro puede ser una distancia entre un átomo de nitrógeno en el cofactor y un átomo de oxígeno en un residuo de tirosina en un sitio de unión, otro parámetro puede ser una distancia entre un carbono carbonílico en un sustrato y un átomo de fósforo en el cofactor, y así sucesivamente.Cada una de estas distancias puede establecerse dentro de umbrales arbitrarios (por ejemplo, la primera distancia puede necesitar menos de 5 angstroms y la segunda distancia puede ser menos de 7,5 angstroms).

[**0136]** La función de adecuación del algoritmo 173 es la precisión predictiva de los modelos de actividad de secuencia 175 entrenados usando diferentes combinaciones de restricciones de parámetros. De esta manera, se evalúa la capacidad de varias combinaciones de parámetros de geometría restringida para entrenar modelos 175 de actividad de secuencia precisos. En ciertas realizaciones, los modelos de actividad de secuencia se entrenan usando máquinas de vectores de soporte.

[**0137]** Las variantes que no están seleccionadas por algoritmo genético 173 se eliminan de la consideración para producir el conjunto de datos filtrados por variante 177. En otras palabras, el resultado de filtrar por algoritmo genético único 173 es un subconjunto de los datos en bruto 171 que contiene sólo los datos para un subconjunto de las variantes en los datos 171. Este subconjunto se usa para entrenar un modelo de actividad de secuencia altamente preciso que a su vez se usa en otro algoritmo genético, un algoritmo de predicción 179. En ciertas realizaciones, el algoritmo de predicción 179 identifica nuevas secuencias variantes predichas para tener una alta actividad. Puede hacer esto aplicando secuencias alternativas de aminoácidos (o nucleótidos) al modelo de actividad de secuencia entrenado y determinando cuáles tienen valores altos para una propiedad beneficiosa (por ejemplo, la actividad del modelo de actividad de secuencia). El algoritmo genético 179 genera secuencias alternativas, las cuales son evaluadas por el modelo de actividad de secuencia entrenada para la aptitud. En última instancia, las secuencias variantes de alto rendimiento se identifican para una mayor investigación y/o producción.

III. DESCRIPCIÓN GENERAL DE APLICACIONES DE ALGORITMO GENÉTICO

[0138] Algunas realizaciones proporcionan métodos de uso de los algoritmos genéticos para generar un conjunto de datos filtrados para la formación de un modelo de actividad de la secuencia tal como uno optimizado por una máquina de vectores de soporte (por ejemplo, el primer y segundo algoritmos genéticos descritos a continuación). Otras realizaciones proporcionan métodos para usar algoritmos genéticos para ajustar los valores de los coeficientes de los modelos de actividad de secuencia para ajustar los modelos al conjunto de datos de entrenamiento filtrado. Sin embargo, otras realizaciones usan un algoritmo genético para explorar el espacio de secuencia e identificar variantes de proteínas que tienen propiedades ventajosas (por ejemplo, el tercer algoritmo genético descrito a continuación).

[**0139]** En un algoritmo genético, se definen una función física adecuada y un procedimiento de apareamiento apropiado. La función de aptitud proporciona un criterio para determinar qué "individuos" (modelos en algunas realizaciones) están "más en forma" con respecto a los datos observados o tienen el poder predictivo más alto (es decir, es probable que los modelos proporcionen los mejores resultados). En algunas realizaciones, un modelo se define por una relación entre una o más variables independientes (IV) y una variable dependiente (DV), y la relación se describe por uno o más parámetros. El algoritmo genético proporciona un mecanismo para buscar en los espacios de parámetros para encontrar las combinaciones de parámetros o rangos de valores de parámetros que generan los modelos más exitosos.

[**0140]** Muchos procesos en algoritmos genéticos están inspirados en operaciones genéticas biológicas. Como tal, los términos utilizados en algoritmos genéticos se toman prestados de términos biológicos relacionados con operaciones genéticas. En estas realizaciones, cada uno de los "individuos" (a veces denominados miembros o cromosomas) de una población incluye "genes" que representan todos los parámetros que se prueban para un modelo, y los genes que

tienen valores elegidos en rangos definidos para los parámetros. Por ejemplo, un cromosoma puede tener un gen que represente la presencia de Gly en la posición 131.

[**0141]** En algunas realizaciones, el algoritmo genético puede usarse para seleccionar IVs apropiados para los modelos (por ejemplo, el primer algoritmo genético descrito a continuación para el filtrado de columnas). Un ejemplo de dicho algoritmo incluye genes/parámetros de valor binario 1 y 0, cada parámetro asociado con un IV. Si un parámetro converge a 0 para un IV entre los individuos más aptos al final del algoritmo, ese IV se elimina del modelo. Ese término se conserva a la inversa.

[**0142]** En algunas realizaciones, la aptitud de un modelo se mide por el poder predictivo del modelo. En algunas realizaciones del documento, la aptitud se mide mediante tasas de aciertos basadas en una matriz de confusión descrita a continuación. En algunas realizaciones, la aptitud se mide por AIC o BIC. Los modelos en este ejemplo pueden en algunos casos ser los conjuntos de datos subyacentes utilizados para producir esos modelos.

[**0143]** Después de que cada "modelo" en una generación particular se evalúa por su poder predictivo, se verifica la convergencia del algoritmo genético u otros criterios (como un número fijo de generaciones) para determinar si el proceso debe continuar durante una generación más. Suponiendo que el algoritmo genético aún no ha cumplido el criterio para detenerse, se clasifican los modelos de la generación actual. Aquellos que tienen el mayor poder predictivo pueden ser preservados y utilizados en la próxima generación. Por ejemplo, se puede emplear una tasa de elitismo del 10%. En otras palabras, el 10% superior de los modelos (según lo determinado utilizando la función de ajuste y medido por, por ejemplo, precisión o AIC) se reserva para convertirse en miembros de la próxima generación. El 90% restante de los miembros de la próxima generación se obtienen al aparear "padres" de la generación anterior.

[**0144]** Como se indica, los "padres" son modelos seleccionados de la generación anterior. En general, la selección está ponderada hacia miembros más en forma de la generación anterior, aunque puede haber un componente aleatorio en su selección. Por ejemplo, los modelos primarios pueden seleccionarse usando una ponderación lineal (por ejemplo, un modelo que rinde 1,2 veces mejor que otro modelo tiene un 20% más de probabilidades de ser seleccionado) o una ponderación geométrica (es decir, las diferencias predictivas en los modelos se elevan a un poder para obtener una probabilidad de selección). En algunas realizaciones, los padres se seleccionan simplemente eligiendo los dos o más modelos de mejor rendimiento de la clasificación de modelos de la generación anterior y no se seleccionan otros modelos. En estas realizaciones, todos los modelos seleccionados de la generación anterior están acoplados. En otras realizaciones, algunos modelos de la generación anterior se seleccionan para su inclusión en el modelo de la próxima generación sin apareamiento, y otros modelos de peor rendimiento de la generación anterior se seleccionan aleatoriamente como padres. Estos padres pueden emparejarse entre sí y/o con los modelos de mejor rendimiento seleccionados para su inclusión como tales en la próxima generación.

[**0145]** Después de que se ha seleccionado un conjunto de modelos principales, los pares de tales modelos se combinan para producir modelos secundarios proporcionando algunos genes (valores de parámetros) de un padre y otros genes (valores de parámetros) del otro padre. En un enfoque, los coeficientes de los dos padres están alineados y cada valor se considera sucesivamente para determinar si el niño debe tomar el término del padre A o del padre B. En una implementación, el proceso de apareamiento comienza con el padre A y determina aleatoriamente si debe ocurrir un evento "cruzado" en el primer término encontrado. Si es así, el término se toma del padre B. Si no, el término se toma del padre A. El siguiente término en sucesión se considera cruzado, etc. Los términos continúan viniendo del padre que dona el término anterior bajo consideración hasta se produce un evento cruzado. En ese punto, el próximo término se dona del otro padre y todos los términos sucesivos se donan de ese padre hasta que ocurra otro evento cruzado. Para garantizar que no se seleccione el mismo término en dos ubicaciones diferentes en el modelo secundario, se pueden emplear varias técnicas, por ejemplo, una técnica cruzada parcialmente coincidente. En algunas realizaciones, en lugar de usar los valores de los genes de cualquiera de los padres, el promedio de los valores del gen puede adoptarse para un cromosoma infantil.

[**0146]** En algunas realizaciones, un algoritmo genético también emplea uno o más mecanismos de mutación para generar mayor diversidad de los modelos, lo que ayuda a explorar regiones de un espacio de parámetros que no están cubiertos por ningún gen existente en la generación parental. Por otro lado, los mecanismos de mutación afectan la convergencia, de modo que cuanto mayor sea la tasa de mutación o mayor el rango de mutación, más tardará en converger (si alguna vez). En algunas realizaciones, la mutación se implementa mediante la selección aleatoria de un cromosoma/modelo, y una selección aleatoria de un parámetro/gen de dicho cromosoma, que luego se cambia aleatoriamente. En algunas realizaciones, los valores cambiados aleatoriamente de parámetros/genes se extraen de una distribución uniforme aleatoria con un rango definido. En otras realizaciones, los valores cambiados aleatoriamente de parámetros/genes se extraen de una distribución normal aleatoria con un rango definido.

[**0147]** Después de considerar cada parámetro, se define un "modelo" hijo para la próxima generación. Luego, se pueden elegir otros dos padres para producir otro modelo infantil, y así sucesivamente. Finalmente, la población infantil de una nueva generación está lista para ser evaluada por la función de forma física descrita anteriormente.

[**0148]** El proceso continúa generación por generación hasta cumplir un criterio de detención, como la convergencia de valores. En ese momento, al menos uno de los modelos mejor clasificados se selecciona de la generación actual

como el mejor modelo general. La convergencia puede ser probada por muchas técnicas convencionales. En algunas realizaciones, implica determinar que el rendimiento del mejor modelo de varias generaciones sucesivas no cambia apreciablemente. Los ejemplos de criterios de detención incluyen, entre otros, el número de generaciones generadas hasta el momento, la actividad de las proteínas superiores de la biblioteca actual, la magnitud de la actividad deseada y el nivel de mejora observado en la última generación de modelos.

IV. REALIZACIONES QUE USAN ALGORITMOS GENÉTICOS PARA EL FILTRADO DE DATOS

[0149] En algunas realizaciones, hay dos o tres etapas para obtener y usar un modelo de actividad de secuencia a partir de la información disponible. Cada uno de estos pasos utiliza un algoritmo genético. En un proceso de tres etapas, un primer algoritmo genético opera con datos de un conjunto de datos sin procesar para seleccionar variables independientes para su uso en un modelo de actividad de secuencia. Estas variables independientes se seleccionan del conjunto de variables independientes disponibles (a veces llamadas parámetros). No todas las variables independientes disponibles se utilizan en el modelo final. En una realización, la información de secuencia o mutación siempre se usa como una variable independiente, pero un algoritmo genético selecciona otros tipos de variables independientes. Se selecciona una combinación particular de variables independientes que hace un muy buen trabajo (o, en algunas realizaciones, el mejor trabajo) para predecir con precisión la actividad. Como ejemplo, puede haber de cinco a diez variables independientes disponibles para usar además de la información de secuencia, pero solo tres de estas variables que no son de secuencia se seleccionan para usar en un modelo de actividad de secuencia. Un algoritmo genético identifica cuál de las muchas combinaciones alternativas de variables independientes hace el mejor trabajo al entrenar un modelo de actividad de secuencia para predecir la actividad.

[**0150]** Otro algoritmo genético identifica rangos adecuados de algunas o todas las variables independientes de no secuencia en el conjunto de datos. Los rangos pueden definirse por umbrales o valores de corte para las variables independientes. Este algoritmo genético se utiliza en procesos de dos y tres etapas.

[**0151]** Un algoritmo genético final identifica secuencias de biomolécula (por ejemplo, variante de proteína) que merecen selección o análisis adicional. Este algoritmo genético proporciona varias secuencias y prueba su aptitud utilizando un modelo de actividad de secuencia entrenado usando datos filtrados seleccionados usando uno o dos algoritmos genéticos anteriores. Vale la pena notar una diferencia entre este algoritmo genético y otros algoritmos genéticos discutidos aquí. Este algoritmo proporciona ácido nucleico, aminoácidos u otras secuencias de biomoléculas como individuos en una población. En contraste, en otro algoritmo genético discutido aquí, los individuos son modelos o conjuntos de parámetros modelo.

[**0152]** En algunas realizaciones, el modelo de actividad de secuencia es un modelo no lineal. En otras realizaciones, es un modelo lineal.

[**0153]** Como se ilustra en la Figura 2, los datos disponibles para un conjunto de entrenamiento del modelo de actividad de secuencia incluyen información para cada una de las múltiples variantes biomoleculares utilizadas para preparar el conjunto de entrenamiento. La información para cada variante incluye su secuencia y su actividad. En varios ejemplos presentados en este documento, la actividad es la velocidad y/o la estereoselectividad de una biomolécula enzimática al volcar un sustrato. Se pueden emplear otros tipos de actividad o propiedades beneficiosas y algunos de estos tipos se describen en otra parte del presente documento. Los datos de la actividad se determinan a partir del análisis *in vitro* y/o una técnica de cálculo como el cribado virtual.

[**0154]** En ciertas realizaciones, la información de secuencia puede proporcionarse como un grupo de mutaciones a una cadena principal inicial, la cual puede ser una secuencia de tipo salvaje o alguna otra secuencia tal como una secuencia consenso. La información de secuencia con respecto a las mutaciones puede presentarse en forma del residuo inicial y el residuo sustituto en una posición dada. Otra alternativa simplemente identifica el residuo final en una posición particular. En diversas realizaciones, la información de secuencia se proporciona mediante un algoritmo genético u otra técnica computacional y, por lo tanto, se conoce sin la necesidad de secuenciar un ácido nucleico u otra composición. Si se requiere secuenciación, se puede emplear cualquiera de los muchos tipos de secuenciación. Algunos de estos tipos se describen en otra parte del presente documento. Por ejemplo, en algunas realizaciones, se usan técnicas de alto rendimiento para secuenciar ácidos nucleicos.

[**0155]** Además de los datos de secuencia y actividad, los datos en bruto contienen varios tipos de información adicional que pueden incorporarse, o no, en el conjunto de entrenamiento final para el modelo de actividad de secuencia. La información adicional puede ser de muchos tipos diferentes. Cada tipo potencialmente sirve como una variable independiente para un modelo de actividad de secuencia. Como se explica aquí, un algoritmo genético u otra técnica evalúa la utilidad de cada tipo de información.

[**0156]** En diversas realizaciones, la información adicional describe características de la unión ligando-receptor. Dicha información puede derivarse de mediciones y/o cálculos. Como se mencionó, los modelos estructurales de variantes pueden identificar valores para estos otros tipos de información. En un ejemplo, el modelo estructural es un modelo de homología. Se puede utilizar una herramienta acoplable o similar para obtener información adicional del modelo estructural. Los ejemplos de información generada a partir de un acoplador incluyen las energías de interacción y/o

las energías totales calculadas por un programa de acoplamiento como el programa Accelrys CDocker. Otros ejemplos se refieren a parámetros geométricos que caracterizan la posición relativa del ligando o sus restos o átomos activos con respecto a un cofactor, un residuo del sitio de unión y/u otra característica asociada con el sitio de unión de la variante en consideración. Como se mencionó, parte de esta información puede referirse a distancias, ángulos y/o información de torsión sobre las posiciones relativas del sustrato o intermedio y un cofactor o residuo en el sitio de unión. Como ejemplos, los valores de energía de interacción pueden basarse en la fuerza de van der Waals y/o la interacción electrostática. La energía interna del ligando también se puede considerar.

[**0157]** La Figura 2A-2C ilustra un ejemplo de filtrado de un conjunto de datos de actividad de secuencia sin procesar de acuerdo con algunas realizaciones de la divulgación actual. La Figura 2A muestra un conjunto de datos de actividad de secuencia sin procesar para n variantes de una familia de transaminasas. Cada variante está asociada con datos de actividad, datos de secuencia, datos de energía y datos de geometría. En algunas realizaciones, los datos de actividad pueden ser velocidad catalítica, enantioespecificidad, etc., que pueden analizarse mediante diversos métodos descritos en otra parte del presente documento. Se proporcionan tres posiciones de secuencia para cada variante, P1, P2 y P3, en el conjunto de datos sin procesar para su inclusión en el modelo de actividad de secuencia. Además, se proporcionan dos valores de energía, la energía total y la energía de interacción determinada por un sistema de acoplamiento virtual en otra parte del presente documento para su posible inclusión en el modelo. Finalmente, un sistema de acoplamiento virtual proporciona cinco valores de geometría para una posible inclusión en el modelo. En este ejemplo que involucra un ligando, cada uno de estos valores de geometría es la distancia entre un átomo clave del ligando cuando está acoplado a la variante enzimática versus cuando está acoplado a la enzima de tipo salvaje. Específicamente, N_1 denota un átomo de nitrógeno, P es un fósforo de un grupo fosfato, C(o) es el átomo de carbono de un grupo carboxilo, $C_{(H3)}$ es el átomo de carbono de un grupo metilo, y $O_{(H)}$ es el átomo de oxígeno de un grupo hidroxilo.

[**0158]** Según algunas realizaciones, los datos de actividad de secuencia sin procesar se pueden filtrar mediante un algoritmo genético para excluir columnas de datos que no son informativas para entrenar un modelo de actividad de secuencia de alto poder predictivo. La Figura 2B muestra un ejemplo de columnas de datos que se filtran mediante un algoritmo genético. En esta implementación, el algoritmo genético genera una población de individuos, cada individuo con un conjunto de "genes" o coeficientes con valores binarios (p. ej., 0 y 1) que indican si los valores de energía y geometría deben incluirse en el modelo de actividad de secuencia. El ejemplo en la Figura 2B muestra el efecto de un individuo de una población de la AG, el individuo tiene los siguientes parámetros: E Total = 1, E Interactuar = 1, N_1 = 1, P = 1, C(o) = 0, $C_{(H3)}$ = 1, $O_{(H)}$ = 0. Cuando un parámetro toma el valor de 0, la característica asociada con el parámetro se excluye efectivamente del modelo. Esta GA individual filtra los datos de geometría C(o) y $O_{(H)}$, proporcionando así un subconjunto de datos para entrenar un modelo de actividad de secuencia. En algunas realizaciones, se entrena un modelo de actividad de secuencia usando el subconjunto de datos que incluye tres IV de secuencia, dos IV de energía y tres IV de geometría. Tenga en cuenta que los coeficientes de valor binario o genes de la GA pueden implementarse por separado del modelo de actividad de secuencia, de modo que el modelo de actividad de secuencia no incluye los valores de coeficiente. En algunas realizaciones, el modelo de actividad de secuencia se optimiza usando un SVM, que genera aciertos y errores para la actividad predicha. La función de aptitud física de la GA determinada para cada individuo se basa en la precisión de la predicción. Múltiples individuos en la población de una generación de GA se prueban de la misma manera descrita anteriormente. Cada individuo tiene un conjunto de parámetros con valores de 0 o 1, en donde los parámetros con valor 0 filtran efectivamente un conjunto de características, produciendo así un subconjunto de datos para entrenar un modelo de actividad de secuencia. Los individuos se comparan y clasifican según sus funciones físicas. Luego, uno o más de los individuos "más aptos" se seleccionan como padres para una próxima generación de población utilizando al menos un mecanismo de diversidad, como se describe en otra parte del presente documento. En algunas realizaciones, la comparación de la aptitud se implementa usando el Criterio de Información de Akaike (AIC) o el Criterio de Información Bayesiano (BIC), en donde los individuos que tienen los valores de AIC o BIC más pequeños se eligen como los individuos más adecuados. Por lo general, la AG se repite durante dos o más generaciones hasta que se cumpla un criterio de convergencia.

[**0159]** Tenga en cuenta que el filtrado de columnas es opcional en algunas realizaciones. Según algunas realizaciones, los datos de actividad de secuencia sin procesar se pueden filtrar mediante un algoritmo genético para excluir filas de datos en lugar de o además del filtrado de columnas. La Figura 2C muestra un ejemplo de filas de datos (variantes enzimáticas) que se filtran mediante un algoritmo genético. En esta implementación, el algoritmo genético proporciona una población de individuos, cada individuo con un conjunto de "genes" o coeficientes de valor continuo que indican un valor umbral de exclusión. Si los valores de energía y geometría están por encima del umbral para una variante, la variante se excluye del modelo de actividad de secuencia. El ejemplo en la Figura 2C muestra un individuo GA que tiene los siguientes valores de umbral: E Total > 1,5, E Interacción > 1,5, N_1 > 3,3, P > 2,8, C(o) > 3,6, $C_{(H3)}$ > 6 y $O_{(H)}$ > 6. Estos valores de umbral son solo para fines ilustrativos y no indican umbrales óptimos para implementaciones reales. En este ejemplo, este GA individual filtra la variante 1 y la variante 5, proporcionando un subconjunto de datos para entrenar el modelo de actividad de secuencia. Tenga en cuenta que los valores de umbral de GA pueden implementarse por separado del modelo de actividad de secuencia, de modo que el modelo de actividad de secuencia no incluye los valores de umbral. Como en el filtrado de columnas, en algunas realizaciones, el modelo de actividad de secuencia se optimiza usando un SVM, que genera aciertos y errores para la actividad prevista. La función de aptitud del individuo se basa en la precisión de la predicción. Múltiples individuos de la GA se prueban de la misma manera descrita en el ejemplo anterior. Los individuos se comparan y clasifican según sus funciones físicas.

Luego, se seleccionan uno o más individuos más aptos para generar una próxima generación de población utilizando al menos un mecanismo de diversidad, como se describe en otra parte del presente documento.

[**0160]** En algunas realizaciones, los individuos más aptos derivados de la GA mostrados en los ejemplos de la Figura 2 proporcionan subconjuntos de datos y entrenan una máquina de vectores de soporte, para definir los parámetros de un modelo de actividad de secuencia que tiene un alto poder predictivo. En algunas realizaciones, este modelo de actividad de secuencia puede guiar el diseño de nuevas variantes para una nueva ronda de evolución dirigida, como se describe más adelante. Después de obtener uno o más de los "mejores modelos de actividad de secuencia", algunas realizaciones usan estos modelos para guiar la síntesis de proteínas reales, que pueden desarrollarse adicionalmente por evolución dirigida. Algunas realizaciones proporcionan métodos para diseñar proteínas con la actividad deseada modificando secuencias predichas por el modelo, como se describe en otra parte del presente documento.

A. Primer algoritmo genético: selección de parámetros

[0161] En ciertas realizaciones, tales como la realización representada en la Figura 3A, un algoritmo genético selecciona parámetros particulares del conjunto de parámetros disponibles, así como la información de actividad para múltiples variantes. La realización mostrada en la Figura 3A es una forma de implementar el paso 105 de filtrado de datos sin procesar para eliminar uno o más tipos de energía y/o tipos de geometría en el proceso representado en la Figura 1A. Los datos para estos parámetros se proporcionan en un conjunto de datos sin filtrar. Ver el bloque 303 de la Figura 3A. Todos los datos pueden combinarse en uno o más archivos legibles por computadora para un acceso conveniente durante la ejecución del primer algoritmo genético.

[**0162]** Para implementar el primer algoritmo genético, se usa un grupo de parámetros seleccionados al azar del conjunto de parámetros disponibles para proporcionar una primera generación de subconjuntos de datos. Ver bloque 305. Cada colección de parámetros, que sirven como colecciones de variables independientes, define un subconjunto de datos único. Los diferentes grupos de variables independientes seleccionados al azar (es decir, múltiples subconjuntos de datos individuales) se utilizan para entrenar los modelos de actividad de secuencia. En algunas realizaciones, se usa el mismo número de variables independientes para crear cada subconjunto de datos. En muchas implementaciones, la información de secuencia o mutación se usa como una variable independiente adicional en cada subconjunto de datos. Colectivamente, los subconjuntos de datos conforman los "individuos" en una población de una generación de un algoritmo genético.

[**0163]** En la primera generación del algoritmo genético, se proporcionan modelos de actividad de secuencia de cada uno de los subconjuntos de datos con cada modelo asociado con una combinación diferente seleccionada aleatoriamente de variables independientes. Estos se utilizan para predecir la actividad. Ver bloque 307. En ciertas realizaciones, la predicción se realiza en secuencias que no se usaron para entrenar realmente el modelo, probando el poder predictivo del modelo mediante validación cruzada. Por ejemplo, los datos sin filtrar pueden estar disponibles para 100 variantes, pero los datos de solo 70 de estos se utilizan para entrenar los modelos de actividad de secuencia. Las 30 variantes restantes, o más precisamente los datos para estas 30 variantes restantes, se usan como un conjunto de prueba para probar la efectividad de los modelos de actividad de secuencia, proporcionando validación cruzada del poder predictivo del modelo.

[**0164]** Los subconjuntos de datos resultantes obtenidos durante esta primera generación del primer algoritmo genético se clasifican en función de su capacidad para entrenar modelos que predicen con precisión la actividad. Vea el bloque 311. La clasificación se realiza utilizando una función de aptitud física que puede verse como el rendimiento de los modelos entrenados. En otras palabras, el proceso deriva modelos de los datos sin procesar filtrados de diferentes maneras para eliminar diferentes combinaciones de variables. Los modelos evalúan la idoneidad de los subconjuntos de datos (es decir, individuos) que se usaron para entrenarlos.

[**0165]** Los subconjuntos de datos clasificados más bajos reflejan las colecciones clasificadas más bajas de variables independientes y se rechazan antes de pasar a la segunda generación del algoritmo genético. Los subconjuntos de datos rechazados se reemplazan con subconjuntos de datos derivados de los tipos de modelos de alto rendimiento de la primera generación. Ver bloque 313.

[**0166] El** acoplamiento de subconjuntos de datos puede realizarse mediante diversas técnicas. Básicamente, algunas de las variables independientes seleccionadas de cada uno de los dos subconjuntos de datos parentales se utilizan en el apareamiento, por lo que pueden trasladarse al subconjunto de datos secundarios. En un ejemplo, dos subconjuntos de datos primarios se representan como una secuencia de 1s y 0s para indicar si los parámetros particulares del conjunto de variables independientes disponibles se usan como variables independientes en los subconjuntos de datos. Estas representaciones binarias de los subconjuntos de datos se cortan en un punto cruzado y los segmentos resultantes se unen con segmentos complementarios del subconjunto de datos del otro padre.

[**0167]** La función de aptitud, o más precisamente el método de evaluación de la precisión de un modelo de actividad de secuencia particular, puede implementarse de varias maneras. En un enfoque, la función de aptitud evalúa la precisión del modelo utilizando una matriz de confusión. En dicha técnica, cada una de las variantes utilizadas en un

conjunto de prueba se considera activa o inactiva, dependiendo de si su actividad medida es mayor o menor que un umbral definido. Similarmente, el modelo de actividad de secuencia se caracteriza por predecir que una variante del conjunto de prueba sea activa o inactiva en función de si predice que el valor de la actividad estará por encima o por debajo del valor umbral definido. Para cada miembro del conjunto de prueba, se comparan los estados de actividad reales y pronosticados del miembro. Un modelo de actividad de secuencia obtiene crédito cuando caracteriza correctamente una variante de prueba como activa o inactiva. Pierde crédito cuando predice que una variante de prueba está inactiva cuando se mide que está activa o cuando predice que una variante de prueba está activa cuando se mide que está inactiva. Estas cuatro alternativas constituyen la matriz de confusión. La frecuencia con la que un modelo particular predice correctamente la actividad o la inactividad se usa para clasificar el subconjunto de datos utilizado para entrenar el modelo. Otra opción para caracterizar la precisión del modelo se basa en el error o la diferencia entre su actividad prevista (o la magnitud del mismo) y la actividad real medida. Esta distancia se puede sumar o promediar sobre todos los miembros del conjunto de prueba.

[**0168]** Al final del algoritmo genético de primera generación, se seleccionan algunos grupos variables independientes (es decir, subconjuntos de datos) para modelos de actividad de secuencia. Como se mencionó, los subconjuntos de datos altamente clasificados se seleccionan para el apareamiento y/o promoción a la próxima generación. Estos subconjuntos contienen estructuras seleccionadas (por ejemplo, distancia) y/o variables independientes de la energía además de la variable independiente de la secuencia.

[**0169]** La segunda generación de subconjuntos de datos se evalúa por la capacidad predictiva de los modelos entrenados para usarlos. El proceso se repite durante varias generaciones hasta que converge la selección de variables independientes. Véase el bloque de convergencia 309. En ciertas realizaciones, un criterio de convergencia determina si la mejora de una generación actual, en comparación con la generación anterior, es menor que un nivel umbral para una o más generaciones consecutivas. En algunas realizaciones, otras formas de probar la convergencia incluyen, pero no se limitan a, probar un valor de condición física máximo o mínimo como 100% de condición física, correr durante un número fijo de generaciones, correr dentro de un límite de tiempo fijo o una combinación de encima. En ciertas realizaciones, se producen y evalúan aproximadamente 5-100 subconjuntos de datos en cada generación. En ciertas realizaciones, se producen y evalúan aproximadamente 30 a 70 subconjuntos de datos en cada generación. No se pretende que la presente invención se limite a ningún número particular de subconjuntos de datos y/o generaciones.

B. Segundo algoritmo genético

[0170] En un segundo algoritmo genético como se ejemplifica en la Figura 3B, se proporciona un proceso para implementar el paso 107 de la Figura 1A para filtrar datos sin procesar, eliminando así datos para variantes que tienen valores de energía y/o valores de geometría fuera de los rangos definidos. En la Figura 3B, las variables independientes identificadas en el primer algoritmo genético son fijas. Las variables independientes no seleccionadas ya no se consideran relevantes, y el segundo algoritmo genético comienza recibiendo el conjunto de datos filtrado por el primer algoritmo genético. Vea el bloque 323. Se puede suponer que las variables independientes seleccionadas por el primer algoritmo genético son las que probablemente tengan el mayor valor para predecir con precisión la actividad, al menos utilizando la forma del modelo de actividad de secuencia en consideración (por ejemplo, un plano n-dimensional generado por una máquina de vectores de soporte). En realizaciones alternativas, el primer algoritmo genético no se realiza y se utilizan todas las variables independientes del conjunto de datos sin procesar.

[**0171]** Debe entenderse que las secuencias de las variantes establecen necesariamente los valores de las variables independientes adicionales: las variables de restricción energética y estructural. Por ejemplo, la combinación de mutaciones presentes en el bolsillo de unión definirá ciertas características de unión estructural geométrica y los valores de energía de interacción que sirven como variables independientes disponibles. Sin embargo, la información de secuencia por sí sola puede ser inadecuada para entrenar eficazmente el modelo de actividad de secuencia para predecir con precisión la actividad.

[**0172]** En el segundo algoritmo genético, cada variable independiente (que no sea la secuencia) se refina de tal manera que solo las variantes que cumplen un valor umbral de una variable independiente se seleccionan para su uso en el subconjunto de datos. Este refinamiento puede aplicarse a múltiples variables independientes no secuenciales. En otras palabras, el segundo algoritmo genético selecciona un subrango dentro del rango total disponible de magnitudes para una o más de las variables independientes no secuenciales seleccionadas. Como ejemplo de un enfoque, una variable independiente dada puede tener un rango dinámico de aproximadamente 0 a 20 Å, que representa la distancia entre dos átomos o entre dos posiciones acopladas del mismo átomo. Una versión más refinada de esta variable independiente considera solo variantes que tienen valores de aproximadamente 12 Å o menos. Otro ejemplo de un rango de valores puede ser de aproximadamente 5 Å o menos. Un objetivo del segundo algoritmo genético es centrarse en la porción del rango completo de magnitudes variables que son útiles para predecir la actividad. Esto parece reducir el ruido en la capacidad predictiva de los modelos entrenados.

[**0173]** En la primera generación de este segundo tipo de algoritmo genético, cada una de las variables independientes (aparte de la variable de secuencia) se divide en una porción. La partición se realiza al azar. Ver bloque 325. Por ejemplo, se seleccionan aleatoriamente valores particulares de magnitud para cada una de las variables

independientes. Solo se consideran las variantes que tienen valores inferiores a este punto de partición. Esto efectivamente empareja las variables independientes utilizadas en el conjunto de entrenamiento para el modelo de actividad de secuencia.

[**0174]** En la primera generación, los subconjuntos de datos individuales tienen puntos de corte seleccionados al azar para cada variable independiente de la secuencia. Bloque 325. Cada subconjunto de datos individuales en los trenes de primera generación usando su propio modelo único de actividad de secuencia. Consulte el bloque 327. Los modelos resultantes se usan para predecir la actividad de cada miembro de un conjunto de pruebas. Bloque 327. Cada subconjunto de datos individuales se clasifica por su capacidad para entrenar un modelo preciso mediante, por ejemplo, una matriz de confusión como se describió anteriormente. Ver bloque 331. Esta es la función de aptitud. Las funciones alternativas de fitness son posibles. Estas incluyen funciones que utilizan valores de diferencia entre el valor predicho y el real. La aptitud también se puede basar en los tipos de variables independientes utilizadas en los modelos y/o la fracción del rango completo de valores de variables independientes utilizados.

[**0175]** En ciertas realizaciones, un subconjunto de datos contiene datos para un subconjunto de las variantes en el conjunto de datos sin procesar. Los datos para una fracción de estas variantes se utilizan para entrenar un modelo de actividad de secuencia. Los datos para las variantes restantes se utilizan para probar el modelo de actividad de secuencia resultante. En otras palabras, cada subconjunto de datos se divide en un conjunto de entrenamiento y un conjunto de prueba. La división puede realizarse por selección aleatoria. En algunas realizaciones, el conjunto de entrenamiento contiene entre aproximadamente 20 y 90% (o entre aproximadamente 50 y 80%) de las variantes en el subconjunto. No se pretende que la presente invención se limite a ningún número particular de variantes en los subconjuntos y/o conjuntos de entrenamiento.

[**0176]** Los subconjuntos de datos de alta puntuación en la primera generación se seleccionan para uso en la segunda generación y/o como padres para el apareamiento para producir descendencia para la segunda generación. Vea el bloque 333. El apareamiento puede llevarse a cabo utilizando cualquier técnica adecuada. En una realización, se aplica un esquema de ponderación de costos, tal como una suma ponderada de diferencias usando los valores de corte (es decir, umbral) para cada uno de los dos padres de apareamiento para una variable independiente dada. En un esquema de ponderación de costos, la selección de apareamiento está sesgada hacia individuos (es decir, subconjuntos de datos) que tienen una aptitud relativamente mayor. Los individuos más aptos se aparean más que los individuos menos aptos. Otros esquemas de selección de apareamiento incluyen selección de rueda de ruleta proporcional, selección de rueda de ruleta basada en rango y selección de torneo.

[**0177]** El proceso de apareamiento real puede tomar muchas formas. Un ejemplo es el apareamiento continuo de parámetros. En este enfoque, el valor de corte para un parámetro dado en un subconjunto de datos secundarios es un valor que se encuentra entre los valores de corte para el mismo parámetro en los dos subconjuntos de datos principales. Por ejemplo, un padre puede tener un valor de corte de 0,1 angstroms para un primer parámetro (distancia X), mientras que el otro padre puede tener un valor de corte de 0,6 angstroms para la distancia X. El valor de corte del niño para la distancia X estará entre 0,1 y 0,6 angstroms. Se pueden definir varias funciones para determinar el valor de corte intermedio del niño para la distancia X. En un esquema de apareamiento de parámetros continuos, se elige aleatoriamente un valor "beta" y se aplica para determinar la distancia fraccional entre los dos valores de corte de los padres. En el ejemplo anterior, si beta es elegido para ser 0,7 y dos hijos son producidos, los valores de corte de los niños se pueden calcular como sigue:

distancia de niño 1 = 0,1 – (0,7) * 0,1 + (0,7) * 0,6 = 0,45

distancia de niño 2 = 0,6 – (0,7) * 0,1 + (0,7) * 0,6 = 0,25

niño 1 = a + beta * (b – a)

niño 2 = b + beta * (a – b)

[0178] En una segunda generación, los individuos (subconjuntos de datos definidos) seleccionados y/o producidos por apareamiento en la primera ronda se evalúan aplicando la función de aptitud física a cada uno de ellos. En otras palabras, el proceso de los bloques 327, 331 y 333 se aplica a la segunda generación. Al igual que con la primera generación, los subconjuntos de datos pueden clasificarse en función de su capacidad para entrenar modelos que predicen con precisión la actividad en un conjunto de variantes de prueba. Los subconjuntos de alto rango pueden pasar a la siguiente generación y/o aparearse como se describe anteriormente.

[**0179]** Otras generaciones, continúan como con la segunda generación hasta que se alcanza la convergencia. Como se muestra en la Figura 3B, cada generación está sujeta a una verificación de convergencia. Ver bloque 329. En ciertas realizaciones, un criterio de convergencia determina si la mejora de la generación actual, en comparación con la generación anterior, es menor que un nivel umbral para una o más generaciones consecutivas. Otras formas de probar la convergencia incluyen la prueba de un valor de condición física máxima/mínima, como la condición física al 100%, correr durante un número fijo de generaciones, correr dentro de un límite de tiempo fijo o una combinación de lo anterior.

[**0180]** En ciertas realizaciones, se producen y evalúan aproximadamente 5-100 subconjuntos de datos para cada generación. En ciertas realizaciones, se producen y evalúan aproximadamente 30 a 70 subconjuntos de datos para cada generación. En un ejemplo particular, hay aproximadamente 45 subconjuntos de datos individuales en cada generación del segundo algoritmo genético. Sin embargo, no se pretende que la presente invención se limite a ningún número particular de subconjuntos de datos caracterizados y/o utilizados para cada generación.

[**0181]** En algunos aspectos, este proceso de filtrado del conjunto de datos puede caracterizarse como sigue. Inicialmente, un sistema utiliza un conjunto de datos sin filtrar para crear una población de subconjuntos de datos. Cada uno de estos subconjuntos es un "individuo" en una población de una generación de un algoritmo genético. Cada subconjunto de datos se identifica utilizando umbrales de valores de parámetros (puntos de corte) para parámetros geométricos que caracterizan la unión de un ligando a un sitio de unión de una biomolécula. Cuando el sistema aplica los umbrales de valor de los parámetros, elimina de manera efectiva ciertas variantes del conjunto de datos sin filtrar. En otras palabras, cada subconjunto de datos contiene datos para solo algunas de las variantes incluidas en el conjunto de datos sin filtrar.

[**0182]** Para cada subconjunto de datos (es decir, individual), el sistema divide las variantes constituyentes en las que pertenecen a un conjunto de entrenamiento y las que pertenecen a un conjunto de prueba. Las variantes que pertenecen al conjunto de entrenamiento se usan para entrenar un modelo de actividad de secuencia. El entrenamiento puede llevarse a cabo utilizando una técnica como una máquina de vectores de soporte o mínimos cuadrados parciales. El modelo de actividad de secuencia entrenada resultante se aplica a las variantes del conjunto de prueba. El modelo predice la actividad para cada variante del conjunto de pruebas y, por lo tanto, el sistema evalúa la precisión del modelo de actividad de secuencia y, por lo tanto, su subconjunto de datos asociado. La precisión de cada subconjunto de datos (es decir, individual) en la población de la generación de un algoritmo genético se evalúa de la misma manera.

[**0183]** Para una generación dada de un algoritmo genético, cada uno de los subconjuntos de datos y modelos de actividad de secuencia asociados se clasifican en función de su capacidad para predecir con precisión la actividad de las variantes en el conjunto de prueba asociado. Dentro de la generación, el proceso selecciona los subconjuntos mejor clasificados para su promoción a la próxima generación. Además, el proceso combina algunos de los subconjuntos mejor clasificados para producir subconjuntos secundarios, que también se proporcionan a la próxima generación. Los subconjuntos de datos de próxima generación (es decir, individuos) se tratan como se describe anteriormente. Se tratan y evalúan varias generaciones hasta que se alcanza la convergencia.

C. *Tercer algoritmo genético*

[0184] En el flujo de trabajo descrito, un subconjunto de datos seleccionado filtrando la secuencia sin procesar, la actividad y los datos de estructura entrena un modelo de actividad de secuencia de alta precisión. Se puede utilizar una máquina de vectores de soporte para realizar el entrenamiento. El modelo de actividad de secuencia resultante identifica nuevas variantes biomoleculares. En algunas realizaciones, estas nuevas variantes biomoleculares se usan en al menos una ronda de evolución dirigida. En ciertas realizaciones, se emplea un algoritmo genético final para identificar las nuevas variantes de biomolécula descritas en el bloque 111 de la Figura 1A. Un ejemplo de un algoritmo genético adecuado se representa en la Figura 3C. Como se muestra allí, el proceso comienza con el modelo de actividad de secuencia seleccionado después de concluir el segundo algoritmo genético. Bloque 353.

[**0185]** Como se señaló anteriormente, existe una diferencia entre este algoritmo genético y otros algoritmos genéticos discutidos aquí. Este algoritmo proporciona ácido nucleico, aminoácidos u otras secuencias de biomoléculas como individuos en una población. En contraste, en otro algoritmo genético discutido aquí, los individuos son modelos o conjuntos de parámetros modelo. En una primera generación de este GA, el algoritmo genético proporciona una población aleatoria de individuos, cada uno de los cuales representa una secuencia de proteína (u otra biomolécula) distinta. Bloque 355. Las proteínas individuales difieren entre sí por mutaciones en las posiciones dadas. En algunas implementaciones, las mutaciones se generan aleatoriamente, al menos en la primera generación. Las mutaciones pueden generarse con respecto a una columna vertebral de proteína única, como la columna vertebral de una proteína de tipo salvaje o una columna vertebral de referencia identificada durante una ronda de evolución dirigida.

[**0186]** Los individuos en la primera generación se clasifican o seleccionan usando una función de aptitud que es el modelo de actividad de secuencia entrenado en el subconjunto de datos obtenido al concluir el segundo algoritmo genético (es decir, el modelo pasado en el bloque 353). Vea los bloques 357 y 359. La información de secuencia de identificación para cada biomolécula individual se ingresa al modelo de actividad de secuencia. Esta información puede ser una lista de mutaciones, opcionalmente identificando los residuos iniciales y finales para cada una de las posiciones donde residen las mutaciones. El modelo actúa sobre esta entrada asignando una actividad prevista a cada individuo. Bloque 357. Las biomoléculas individuales que tienen los valores de actividad mejor clasificados (según lo predicho por el modelo) se seleccionan para aparearse y/o transferirse a la próxima generación. Bloques 359 y 363. Los individuos acoplados proporcionan nuevas combinaciones de mutaciones, y cada nueva combinación es miembro de la próxima generación. En ciertas realizaciones, el apareamiento se logra mediante una operación cruzada. Un ejemplo de una operación cruzada en este algoritmo genético puede entenderse como sigue. El padre 1 tiene mutaciones en

las posiciones 12 y 25, y el padre 2 tiene mutaciones en las posiciones 15 y 30. La primera descendencia puede tener mutaciones en la posición 12 del padre 1 y en la posición 30 del padre 2, y la segunda descendencia tendrá mutaciones en las posiciones 25 del progenitor 1 y posición 12 del progenitor 2.

[**0187]** En algunos casos, algunas de las crías producidas por el apareamiento (por ejemplo, el 20% de ellas) se mutan adicionalmente usando cualquier método adecuado, que incluye pero no se limita a mutaciones puntuales. Dichas mutaciones pueden realizarse al azar.

[**0188]** Se derivan generaciones adicionales de poblaciones de biomoléculas distintas como se describe para la segunda generación. La creación de nuevas generaciones se repite hasta que la actividad predicha por el modelo no mejore significativamente durante un número definido de generaciones. En este punto, se considera que la población de biomoléculas ha convergido a una lista final de individuos clasificados que se identifican por un conjunto de mutaciones y una actividad prevista. Una condición de convergencia se muestra en el bloque 361 en la Figura 3C.

[**0189]** En ciertas realizaciones, las biomoléculas individuales de la lista final se sintetizan y se seleccionan *in vitro.* Además, las biomoléculas individuales pueden analizarse para proporcionar restricciones geométricas u otros datos estructurales y/o energía de interacción mediante el uso de software de acoplamiento u otras herramientas. La secuencia resultante, la actividad y los datos estructurales/energéticos se combinan para servir como entrada al flujo de trabajo para una próxima ronda de evolución dirigida. En otras palabras, las proteínas analizadas después del algoritmo genético proporcionan datos que pueden servir como un nuevo conjunto de entrenamiento para una segunda ronda de análisis. Por lo tanto, el algoritmo genético de filtrado de datos se realiza nuevamente pero con un conjunto de entrenamiento completamente nuevo. En algunas realizaciones, el conjunto de datos y el modelo de actividad de secuencia de una ronda de evolución dirigida no se conservan en la siguiente ronda. Es decir, la siguiente ronda comienza de nuevo, buscando un nuevo conjunto de variables independientes utilizando el nuevo conjunto de datos sin filtrar.

[**0190]** En algunas realizaciones, el modelo de actividad de secuencia empleado en el tercer algoritmo genético se entrena usando parámetros de energía y/o estructurales (geométricos) así como información de secuencia. Sin embargo, en ciertas implementaciones, el algoritmo genético final solo ingresa información de secuencia, no información energética y/o estructural, al modelo. En otras palabras, si bien el modelo se desarrolló utilizando secuencias y variables independientes de energía y/o estructurales, el modelo no recibe las variables independientes de energía y/o estructurales cuando se evalúan nuevas secuencias en el tercer algoritmo genético.

[**0191]** En ciertas realizaciones, se evalúan aproximadamente 10 a 10.000 biomoléculas en cada generación. En ciertas realizaciones, se evalúan aproximadamente 100 a 1000 biomoléculas en cada generación. En un ejemplo particular, hay alrededor de 500 biomoléculas individuales en cada generación del tercer algoritmo genético. No se pretende que la presente invención se limite a ningún número particular de biomoléculas que se evalúen.

[**0192]** En algún momento, se completa el proceso descrito anteriormente y se selecciona una o más variantes de la generación actual para investigación adicional, síntesis, desarrollo, producción, etc. En un ejemplo, se usa una variante de biomolécula seleccionada para sembrar una o más rondas de evolución dirigida *in vitro.* Como ejemplo, una ronda de evolución dirigida *in vitro* puede incluir (i) preparar una pluralidad de oligonucleótidos que contienen o codifican al menos una porción de la variante de proteína seleccionada, y (ii) realizar una ronda de evolución dirigida *in vitro* usando la pluralidad de oligonucleótidos Los oligonucleótidos pueden prepararse mediante síntesis génica, fragmentación de un ácido nucleico que codifica parte o la totalidad de la variante de proteína seleccionada, etc. En ciertas realizaciones, la ronda de evolución dirigida *in vitro* incluye fragmentar y recombinar la pluralidad de oligonucleótidos. En ciertas realizaciones, la ronda de evolución dirigida *in vitro* incluye realizar mutagénesis de saturación en la pluralidad de oligonucleótidos.

V. MODELOS DE ACTIVIDAD DE SECUENCIA

[0193] Los métodos y sistemas descritos aquí proporcionan un modelo de actividad de secuencia de alto poder predictivo. En algunas realizaciones, el modelo de actividad de secuencia es un modelo no lineal. En otras realizaciones, es un modelo lineal. Se describen ejemplos de modelos de actividad de secuencia lineal y no lineal en la Patente de EE.UU. Nº 7.747,391, y la Publicación de Solicitud de Patente de EE.UU. Nº 2005/0084907. En diversas realizaciones descritas en el presente documento, el modelo de actividad de secuencia se implementa como un hiperplano n-dimensional, que puede ser generado por una máquina de vectores de soporte. En la siguiente descripción, cuando un modelo de actividad de secuencia se ejemplifica como una máquina de vectores de soporte generada por hiperplano n-dimensional, se pretende que esta forma o el modelo puedan ser sustituidos por otros tipos de modelos lineales y no lineales, como los modelos de mínimos cuadrados, modelos de mínimos cuadrados parciales, regresión lineal múltiple, regresión de componentes principales, regresión de mínimos cuadrados parciales, máquina de vectores de soporte, red neuronal, regresión lineal bayesiana o bootstrap, y versiones de conjunto de estos.

[**0194]** Como se indicó anteriormente, en algunas realizaciones, un modelo de actividad de secuencia usado con las realizaciones de la presente memoria relaciona información de secuencia de proteína con actividad de proteína. La información de la secuencia de proteínas utilizada por el modelo puede tomar muchas formas. En algunas

realizaciones, es una secuencia completa de los residuos de aminoácidos en una proteína. Sin embargo, en algunas realizaciones, la secuencia de aminoácidos completa es innecesaria. Por ejemplo, en algunas realizaciones, es suficiente proporcionar solo aquellos residuos que se van a variar en un esfuerzo de investigación particular. En algunas realizaciones que implican etapas de investigación posteriores, muchos residuos son fijos y solo quedan por explorar regiones limitadas de espacio de secuencia. En algunas de estas situaciones, es conveniente proporcionar modelos de actividad de secuencia que requieran, como entradas, solo la identificación de esos residuos en las regiones de la proteína donde continúa la exploración. En algunas realizaciones adicionales, los modelos no requieren que se conozcan las identidades exactas de los residuos en las posiciones de residuos de interés. En algunas de tales realizaciones, se identifican una o más propiedades físicas o químicas que caracterizan el aminoácido en una posición de residuo particular. En algunas realizaciones, los parámetros geométricos que describen información estructural, por ejemplo, las distancias entre restos, se incluyen en el modelo. Aunque la información estructural puede implementarse en un modelo estructural, también puede implementarse como parte de un modelo de actividad de secuencia. Alternativamente, la información estructural puede usarse para filtrar datos para seleccionar un subconjunto de datos de actividad de secuencia para entrenar un modelo de actividad de secuencia.

[**0195]** Además, en algunos modelos, se emplean combinaciones de tales propiedades. De hecho, no se pretende que la presente invención se limite a ningún enfoque particular, ya que los modelos encuentran uso en diversas configuraciones de información de secuencia, información estructural de información de actividad y/u otras propiedades físicas (por ejemplo, hidrofobicidad, etc.).

[**0196]** En algunas realizaciones descritas anteriormente, las secuencias de aminoácidos proporcionan información para variables independientes para modelos de actividad de secuencia. En otras realizaciones, las secuencias de ácido nucleico, a diferencia de las secuencias de aminoácidos, proporcionan información para variables independientes. En las últimas realizaciones, los IV que representan la presencia o ausencia de nucleótidos de tipos particulares en posiciones particulares de secuencias de nucleótidos se usan como entrada para el modelo. Las proteínas derivadas de las secuencias de nucleótidos proporcionan datos de actividad como la salida del modelo. Un experto en la materia reconoce que diferentes secuencias de nucleótidos pueden traducirse en la misma secuencia de aminoácidos debido a la degeneración de codones, en donde dos o más codones diferentes (es decir, tríos de nucleótidos) codifican el mismo aminoácido. Por lo tanto, diferentes secuencias de nucleótidos pueden relacionarse potencialmente con la misma proteína y actividad proteica. Sin embargo, un modelo de actividad de secuencia que toma información de secuencia de nucleótidos como entrada y actividad de proteína como salida no necesita preocuparse por dicha degeneración. Prácticamente, la falta de una correspondencia uno a uno entre una entrada y una salida puede introducir ruido en el modelo en algunas realizaciones, pero dicho ruido no niega la utilidad del modelo. En algunas realizaciones, dicho ruido puede incluso mejorar el poder predictivo del modelo, porque, por ejemplo, es menos probable que el modelo se ajuste demasiado a los datos. En algunas realizaciones, los modelos generalmente tratan la actividad como una variable dependiente y los valores de secuencia/residuo como variables independientes. Los datos de actividad pueden obtenerse usando cualquier medio adecuado conocido en la técnica, que incluye, pero no se limita a ensayos y/o pantallas diseñadas apropiadamente para medir las magnitudes de la actividad/actividades de interés. Dichas técnicas son bien conocidas por los expertos en la técnica y no son esenciales para la presente invención. De hecho, los principios para diseñar ensayos o pantallas apropiadas son ampliamente conocidos y conocidos en la técnica. Las técnicas para obtener secuencias de proteínas también son bien conocidas y no son clave para la presente invención. Como se mencionó, se pueden usar tecnologías de secuenciación de próxima generación. En algunas realizaciones, la actividad de interés puede ser la estabilidad de la proteína (por ejemplo, estabilidad térmica). Sin embargo, muchas realizaciones importantes consideran otras actividades tales como actividad catalítica, resistencia a patógenos y/o toxinas, actividad terapéutica, toxicidad y similares. De hecho, no se pretende que la presente invención se limite a ningún método(s) de ensayo/selección particular y/o método(s) de secuenciación, ya que cualquier método adecuado conocido en la técnica encuentra uso en la presente invención.

[**0197]** En diversas realizaciones, la forma del modelo de actividad de secuencia puede variar ampliamente, siempre que proporcione un vehículo para aproximar correctamente la actividad relativa de proteínas basándose en la información de secuencia, según se desee. Los ejemplos de la forma matemática/lógica de los modelos incluyen, entre otros, expresiones matemáticas aditivas, multiplicativas, lineales/no interactivas y no lineales/interactivas de varios órdenes, redes neuronales, árboles/gráficos de clasificación y regresión, enfoques de agrupamiento, particionamiento recursivo, máquinas de vectores de soporte y similares.

[**0198]** Varias técnicas para generar modelos están disponibles y encuentran uso en la presente invención. En algunas realizaciones, las técnicas implican la optimización de modelos o la minimización de errores de modelo. Los ejemplos específicos incluyen, entre otros, mínimos cuadrados parciales, regresión de conjunto, bosque aleatorio y varias otras técnicas de regresión, así como técnicas de redes neuronales, partición recursiva, técnicas de máquina de vectores de soporte, CART (árboles de clasificación y regresión) y/o similar. En general, la técnica debe producir un modelo que pueda distinguir los residuos que tienen un impacto significativo en la actividad de aquellos que no lo hacen. En algunas realizaciones, los modelos también clasifican individual residuos o posiciones de residuos en función de su impacto en la actividad. No se pretende que la presente invención se limite a ninguna técnica específica para generar modelos, ya que cualquier método adecuado conocido en la técnica encuentra uso en la presente invención.

[**0199]** En algunas realizaciones que implican modelos aditivos, los modelos se generan mediante una técnica de

regresión que identifica la covarianza de variables independientes y dependientes en un conjunto de entrenamiento. Varias técnicas de regresión son conocidas y ampliamente utilizadas. Los ejemplos incluyen, entre otros, regresión lineal múltiple (MLR), regresión de componentes principales (PCR) y regresión de mínimos cuadrados parciales (PLS). En algunas realizaciones, los modelos se generan usando técnicas que involucran múltiples constituyentes, que incluyen, entre otros, regresión de conjunto y bosque aleatorio. Estos y otros métodos adecuados encuentran uso en la presente invención. No se pretende que la presente invención se limite a ninguna técnica particular.

[**0200]** MLR es la más básica de estas técnicas. Se usa simplemente para resolver un conjunto de ecuaciones de coeficientes para los miembros de un conjunto de entrenamiento. Cada ecuación se relaciona con la actividad de un miembro del conjunto de entrenamiento (es decir, variables dependientes) con la presencia o ausencia de un residuo particular en una posición particular (es decir, variables independientes). Dependiendo del número de opciones de residuos en el conjunto de entrenamiento, el número de estas ecuaciones puede ser bastante grande.

[**0201]** Al igual que MLR, PLS y PCR generan modelos a partir de ecuaciones que relacionan la actividad de secuencia con valores de residuos. Sin embargo, estas técnicas lo hacen de manera diferente. Primero realizan una transformación de coordenadas para reducir el número de variables independientes. Luego realizan la regresión sobre las variables transformadas. En MLR, existe un número potencialmente muy grande de variables independientes: dos o más para cada posición de residuo que varía dentro del conjunto de entrenamiento. Dado que las proteínas y los péptidos de interés a menudo son bastante grandes y el conjunto de entrenamiento puede proporcionar muchas secuencias diferentes, el número de variables independientes puede volverse muy grande rápidamente. Al reducir el número de variables para enfocarse en aquellas que proporcionan la mayor variación en el conjunto de datos, PLS y PCR generalmente requieren menos muestras y simplifican los pasos involucrados en la generación de modelos.

[**0202]** La PCR es similar a la regresión PLS en que la regresión real se realiza en un número relativamente pequeño de variables latentes obtenidas por transformación coordinada de las variables independientes sin procesar (es decir, valores de residuos). La diferencia entre PLS y PCR es que las variables latentes en PCR se construyen maximizando la covarianza entre las variables independientes (es decir, valores de residuos). En la regresión PLS, las variables latentes se construyen de tal manera que se maximice la covarianza entre las variables independientes y las variables dependientes (es decir, valores de actividad). Partial Least Squares regression is described in Hand, D.J., et al. (2001) Principles of Data Mining (Adaptive Computation and Machine Learning), Boston, MA, MIT Press, and in Geladi, et al. (1986) "Partial Least-Squares Regression: a Tutorial," Analytica Chimica Acta, 198: 1-17.

[**0203]** En PCR y PLS, el resultado directo del análisis de regresión es una expresión de actividad que es función de las variables latentes ponderadas. Esta expresión se puede transformar en una expresión para actividad en función de las variables independientes originales mediante la realización de una transformación de coordenadas que convierta las variables latentes nuevamente en las variables independientes originales.

[**0204]** En esencia, tanto PCR como PLS reducen primero la dimensionalidad de la información contenida en el conjunto de entrenamiento y luego realizan un análisis de regresión en un conjunto de datos transformado, que se ha transformado para producir nuevas variables independientes, pero conserva los valores variables dependientes originales. Las versiones transformadas de los conjuntos de datos pueden dar como resultado unas pocas expresiones para realizar el análisis de regresión. En los protocolos en los que no se ha realizado ninguna reducción de dimensión, debe considerarse cada residuo separado para el que puede haber una variación. Este puede ser un conjunto muy grande de coeficientes (p. ej., coeficientes de 2^N para interacciones bidireccionales, donde N es el número de posiciones de residuos que pueden variar en el conjunto de entrenamiento). En un análisis típico de componentes principales, solo se emplean 3, 4, 5 o 6 componentes principales. Sin embargo, no se pretende que la presente invención se limite a ningún número particular de componentes principales.

[**0205]** La capacidad de las técnicas de aprendizaje automático para ajustar los datos de entrenamiento a menudo se denomina "ajuste del modelo" y en las técnicas de regresión tales como MLR, PCR y PLS, el ajuste del modelo se mide típicamente por la diferencia al cuadrado sumada entre medido y valores predichos Para un conjunto de entrenamiento dado, el ajuste óptimo del modelo se logrará utilizando MLR, con PCR y PLS que a menudo tienen un peor ajuste del modelo (mayor error al cuadrado suma entre mediciones y predicciones). Sin embargo, la principal ventaja de usar técnicas de regresión variable latente como PCR y PLS radica en la capacidad predictiva de tales modelos. Obtener un modelo ajustado con un error de suma al cuadrado muy pequeño de ninguna manera garantiza que el modelo pueda predecir con precisión nuevas muestras no vistas en el conjunto de entrenamiento; de hecho, a menudo es el caso opuesto, particularmente cuando hay muchas variables y solo un pocas observaciones (es decir, muestras). Por lo tanto, las técnicas de regresión variable latente (por ejemplo, PCR, PLS), aunque a menudo tienen peores ajustes del modelo en los datos de entrenamiento, generalmente son más robustas y pueden predecir nuevas muestras fuera del conjunto de entrenamiento con mayor precisión.

[**0206]** Máquinas de vectores soporte (SVMs) también pueden ser usadas para generar modelos utilizados en la presente invención. Como se explicó anteriormente, los SVM toman series de secuencias de entrenamiento que se han clasificado en dos o más grupos en función de la actividad como entradas. Las máquinas de vectores de soporte operan ponderando diferentes miembros de un conjunto de entrenamiento de manera diferente dependiendo de qué tan cerca estén de una interfaz de hiperplano que separa los miembros "activos" e "inactivos" del conjunto de

entrenamiento. Esta técnica requiere que el científico primero decida qué miembros del grupo de entrenamiento colocarán en el grupo "activo" y qué miembros del grupo de entrenamiento colocarán en el grupo "inactivo". En algunas realizaciones, esto se logra eligiendo un valor numérico apropiado para el nivel de actividad que sirve como límite entre los miembros "activos" e "inactivos" del conjunto de entrenamiento. A partir de esta clasificación, la máquina de vectores de soporte genera un vector, W, que puede proporcionar valores de coeficientes para las variables independientes individuales que definen las secuencias de los miembros del grupo activo e inactivo en el conjunto de entrenamiento. Estos coeficientes se pueden usar para "clasificar" los residuos individuales como se describe en otra parte del presente documento. La técnica se utiliza para identificar un hiperplano que maximiza la distancia entre los miembros más cercanos del conjunto de entrenamiento en lados opuestos de ese plano.

VI. ACOPLAMIENTO DE PROTEÍNA

[0207] En algunas realizaciones, un sistema virtual de acoplamiento o cribado de proteínas está configurado para realizar diversas operaciones asociadas con la identificación computacional de variantes de biomoléculas que probablemente tengan una actividad deseable, como catalizar de manera eficiente y selectiva una reacción a una temperatura definida. El sistema virtual de acoplamiento de proteínas puede tomar como entradas representaciones de al menos un ligando destinado a interactuar con las variantes. El sistema puede tomar como otras entradas representaciones de las variantes de biomolécula, o al menos los sitios de unión de estas variantes. Las representaciones pueden contener posiciones tridimensionales de átomos y/o restos de los ligandos y/o variantes. Los modelos de homología son ejemplos de las representaciones de las variantes de biomoléculas. En algunas realizaciones, un sistema virtual de detección de proteínas puede aplicar información de acoplamiento y restricciones de actividad para evaluar el funcionamiento de las variantes.

[**0208]** En ciertas realizaciones, un sistema virtual de acoplamiento y cribado de proteínas determina uno o más valores de energía y uno o más valores de geometría con referencia a las relaciones entre restos en dos moléculas diferentes. En algunas realizaciones, los valores de energía pueden incluir una energía de interacción entre un sustrato y una enzima con el sustrato en una o más posturas acopladas con la enzima. En algunas realizaciones, los valores de energía pueden incluir una energía de acoplamiento total que incluye una energía de interacción y una energía interna de los participantes de la interacción de unión. En algunas realizaciones, los valores de geometría pueden incluir valores de distancia, ángulo o torsión entre restos de dos moléculas. En algunas realizaciones, los valores de geometría incluyen la distancia entre los restos correspondientes en un sustrato nativo y deseado, ambos acoplados a la misma enzima. En otras realizaciones, los valores de geometría incluyen la distancia entre un sustrato y una enzima acoplada entre sí.

[**0209]** Cuando se considera el recambio catalítico de un sustrato como la actividad, el sistema virtual de detección de proteínas se puede configurar para identificar posturas que se sabe que están asociadas con una reacción particular. En algunas realizaciones, esto implica considerar una reacción intermedia o un estado de transición en lugar del sustrato mismo. Además del recambio, las posturas pueden evaluarse para otros tipos de actividad, como la síntesis estereoselectiva de enantiómeros, la unión a un receptor de una biomolécula diana identificada como importante para el descubrimiento de fármacos, etc. En algunos casos, la actividad es la unión covalente irreversible o reversible como la inhibición covalente dirigida (TCI).

[**0210]** En ciertas realizaciones, se ejecuta un protocolo para calcular las energías de unión para evaluar la energía de cada postura activa de una variante. En algunas implementaciones, el protocolo puede considerar la fuerza de van der Waals, la interacción electrostática y la energía de solvatación. La solvatación generalmente no se considera en los cálculos realizados por los acopladores. Varios modelos de solvatación están disponibles para calcular las energías de unión. Estos incluyen, entre otros, dieléctricos dependientes de la distancia, nacimiento generalizado con suma por pares (GenBorn), Nacimiento generalizado con membrana implícita (GBIM), Nacimiento generalizado con integración de volumen molecular (GBMV), nacimiento generalizado con un cambio simple (GBSW), y la ecuación de Poisson Boltzmann con área de superficie no polar (PBSA). Los protocolos para calcular las energías de enlace son diferentes o están separados de los programas acoplables. Generalmente producen resultados que son más precisos que los puntajes de acoplamiento, debido en parte a la inclusión de efectos de solvatación en sus cálculos. En diversas implementaciones, las energías de enlace se calculan solo para posturas que se consideran activas.

A. Modelos estructurales de biomoléculas y sus sitios de unión

[0211] En ciertas realizaciones, un sistema informático proporciona modelos tridimensionales para variantes de proteínas (u otras biomoléculas). Los modelos tridimensionales son representaciones computacionales de algunas o todas las secuencias de longitud completa de las variantes de proteínas. Típicamente, como mínimo, las representaciones de cálculo cubren al menos los sitios de unión de las variantes de proteínas.

[**0212]** Como se describe en el presente documento, los modelos tridimensionales pueden ser modelos de homología preparados usando un sistema informático diseñado apropiadamente. Los modelos tridimensionales emplean una plantilla estructural en la que las variantes de proteínas varían entre sí en sus secuencias de aminoácidos. Generalmente, una plantilla estructural es una estructura previamente resuelta por cristalografía de rayos X o RMN para una secuencia que es homóloga a la secuencia modelo. La calidad del modelo de homología depende de la

identidad de secuencia y la resolución de la plantilla de estructura. En ciertas realizaciones, los modelos tridimensionales pueden almacenarse en una base de datos para su uso según sea necesario para proyectos actuales o futuros.

[**0213]** Los modelos tridimensionales de las variantes de proteínas pueden producirse mediante técnicas distintas al modelado de homología. Un ejemplo es el enhebrado de proteínas, que también requiere una plantilla de estructura. Otro ejemplo es el modelado de proteínas *ab initio* o *de novo* que no requiere una plantilla de estructura y se basa en principios físicos subyacentes. Los ejemplos de técnicas *ab initio* incluyen simulaciones de dinámica molecular y simulaciones usando el paquete de software Rosetta.

[**0214]** En algunas realizaciones, las variantes de proteínas varían entre sí en sus sitios de unión. En algunos casos, los sitios de unión difieren entre sí por al menos una mutación en la secuencia de aminoácidos del sitio de unión. La mutación puede realizarse en una secuencia de proteína de tipo salvaje o alguna otra secuencia de proteína de referencia. En algunos casos, dos o más de las variantes de proteínas comparten la misma secuencia de aminoácidos para el sitio de unión pero difieren en la secuencia de aminoácidos para otra región de la proteína. En algunos casos, dos variantes de proteínas difieren entre sí en al menos aproximadamente 2 aminoácidos, o al menos aproximadamente 3 aminoácidos, o al menos aproximadamente 4 aminoácidos. Sin embargo, no se pretende que la presente invención se limite a ningún número específico de diferencias de aminoácidos entre variantes de proteínas.

[**0215]** En ciertas realizaciones, la pluralidad de variantes incluye miembros de la biblioteca producidos por una o más rondas de evolución dirigida. Las técnicas de generación de diversidad utilizadas en la evolución dirigida incluyen la combinación aleatoria de genes, la mutagénesis dirigida al sitio y similares. Se describen ejemplos de técnicas de evolución dirigida en la Patente de EE.UU. Nº 7.024,312, Publicación de Solicitud de Patente de EE.UU. Nº 2012/0040871, Patente de EE.UU. Nº 7.981,614, WO2013/003290, Solicitud PCT Nº PCT/US2013/030526.

B. Acoplamiento de un ligando a variantes de proteína

[0216] Como se explica en el presente documento, el acoplamiento puede emplearse para identificar energía de interacción y/o parámetros geométricos para usar en modelos de actividad de secuencia de entrenamiento. Típicamente, el acoplamiento se realiza mediante un sistema informático programado apropiadamente que usa una representación computacional de un ligando y representaciones computacionales de los sitios de unión de la pluralidad generada de variantes.

[**0217]** Como ejemplo, un acoplador puede configurarse para realizar algunas o todas las siguientes operaciones:

1. Genere un conjunto de conformaciones de ligando usando dinámica molecular de alta temperatura con semillas aleatorias. El acoplador puede generar tales conformaciones sin tener en cuenta el entorno del ligando. Por lo tanto, el acoplador puede identificar conformaciones favorables considerando solo la tensión interna u otras consideraciones específicas del ligando solo. El número de conformaciones a generar puede establecerse arbitrariamente. En una realización, se generan al menos aproximadamente 10 conformaciones. En otra realización, se generan al menos aproximadamente 20 conformaciones, o al menos aproximadamente 50 conformaciones, o al menos aproximadamente 100 conformaciones. Sin embargo, no se pretende que la presente invención se limite a un número específico de conformaciones.

2. Genere orientaciones aleatorias de las conformaciones traduciendo el centro del ligando a una ubicación específica dentro del sitio activo del receptor y realizando una serie de rotaciones aleatorias. El número de orientaciones para refinar se puede establecer arbitrariamente. En una realización, se generan al menos aproximadamente 10 orientaciones. En otra realización, se generan al menos aproximadamente 20 orientaciones, o al menos aproximadamente 50 orientaciones, o al menos aproximadamente 100 orientaciones. Sin embargo, no se pretende que la presente invención se limite a ningún número específico de orientaciones. En ciertas realizaciones, el acoplador calcula una energía "suavizada" para generar combinaciones adicionales de orientación y conformación. El acoplador calcula la energía suavizada utilizando supuestos físicamente poco realistas sobre la permisibilidad de ciertas orientaciones en un sitio de enlace. Por ejemplo, el acoplador puede suponer que los átomos de ligando y los átomos del sitio de unión pueden ocupar esencialmente el mismo espacio, lo cual es imposible debido a la repulsión de Pauli y las consideraciones estéricas. Esta suposición suavizada se puede implementar, por ejemplo, empleando una forma relajada del potencial de Lennard-Jones al explorar el espacio de conformación. Mediante el uso de un cálculo de energía suavizada, la ventana acoplable permite una exploración más completa de las conformaciones que la disponible utilizando consideraciones de energía físicamente realistas. Si la energía suavizada de una conformación en una orientación particular es menor que un umbral especificado, se mantiene la orientación de conformación. Estas conformaciones de baja energía se retienen como "posturas". En ciertas implementaciones, este proceso continúa hasta que se encuentra un número deseado de posturas de baja energía o se encuentra un número máximo de posturas malas.

3. Sujete cada postura retenida del paso 2 a la dinámica molecular de recocido simulado para refinar la postura. La temperatura se aumenta a un valor alto y luego se enfría a la temperatura diana. El acoplador

puede hacer esto para proporcionar una orientación y/o conformación más realista físicamente que la que proporciona el cálculo de energía suavizada.

4. Realice una minimización final del ligando en el receptor rígido utilizando un potencial no ablandado. Esto proporciona un valor energético más preciso para las posturas retenidas. Sin embargo, el cálculo puede proporcionar solo información parcial sobre las energías de las posturas.

5. Para cada postura final, calcule la energía total (energía de interacción receptor-ligando más tensión interna del ligando) y la energía de interacción sola. El cálculo puede realizarse usando CHARMm. Las posturas se ordenan por energía CHARMm y se conservan las posturas de mayor puntuación (más negativas, por lo tanto favorables para la unión). En algunas realizaciones, el documento este paso (y/o paso 4) elimina las posturas que son energéticamente desfavorables.

[**0218]** La siguiente referencia proporciona un ejemplo del funcionamiento de un acoplador: Wu et al., Detailed Analysis of Grid- Based Molecular Docking: A Case Study of CDOCKER - A CHARMm-Based MD Docking Algorithm, J. Computational Chem., Vol. 24, No. 13, pp 1549-62 (2003).

[**0219]** Un acoplador como el que se describe aquí puede proporcionar información tal como la identidad de variantes para las cuales es improbable el acoplamiento con el sustrato deseado, conjuntos de posturas (un conjunto para cada variante) que pueden considerarse para la actividad y energías de interacción para las posturas en los conjuntos.

***C. Determinar los parámetros geométricos del ligando* acoplado**

[0220] Para una variante de proteína que se acopla con éxito al ligando, los parámetros de unión geométrica pueden identificar una o más posturas activas. Una postura activa es una que cumple una restricción más para que el ligando se una en condiciones definidas (en lugar de condiciones de unión arbitrarias). Si el ligando es un sustrato y la proteína es una enzima, la unión activa puede ser una unión que permite que el sustrato experimente una transformación química catalizada, particularmente una transformación estereoespecífica. En algunas implementaciones, las características de unión geométrica definen posiciones relativas de uno o más átomos en el ligando y uno o más átomos en la proteína y/o cofactor asociado con la proteína.

[**0221]** En algunos casos, los parámetros geométricos se identifican a partir de una o más conformaciones de un sustrato nativo y/o intermedio posterior cuando experimenta una transformación química catalizada por una enzima de tipo salvaje. En ciertas realizaciones, los parámetros geométricos incluyen (i) una distancia entre un resto particular en el sustrato y/o intermedio posterior y un residuo particular o resto residual en el sitio catalítico, (ii) una distancia entre un resto particular en el sustrato y/o intermedio posterior y un cofactor particular en el sitio catalítico, y/o (iii) una distancia entre un resto particular en el sustrato y/o intermedio posterior y un resto particular en un sustrato nativo posicionado idealmente y/o intermedio posterior en el sitio catalítico. Las alternativas a la distancia incluyen ángulos entre enlaces o alineaciones atómicas intercompuestas, posiciones torsionales alrededor de un eje común, etc.

[**0222]** Se puede generar una pluralidad de posturas de la representación computacional del sustrato y/o intermedio posterior con respecto a una representación computacional de la variante de proteína en consideración. La pluralidad de posturas puede generarse mediante diversas técnicas. Los ejemplos generales de tales técnicas incluyen, entre otros, búsquedas torsionales sistemáticas o estocásticas sobre enlaces rotativos, simulaciones de dinámica molecular y algoritmos genéticos diseñados para localizar conformaciones de baja energía. En un ejemplo, las posturas se generan utilizando dinámicas moleculares de alta temperatura, seguidas de rotación aleatoria, refinamiento mediante recocido simulado basado en cuadrícula y/o una minimización final basada en cuadrícula o campo de fuerza para generar una conformación y/u orientación del sustrato y/o intermedio posterior en el sitio catalítico de representación computacional. Algunas de estas operaciones son opcionales, por ejemplo, refinamiento mediante recocido simulado basado en cuadrícula y minimización de campo de fuerza o basada en cuadrícula.

[**0223]** En ciertas realizaciones, el número de posturas considerado es al menos aproximadamente 10, o al menos aproximadamente 20, o al menos aproximadamente 50, o al menos aproximadamente 100, o al menos aproximadamente 200, o al menos aproximadamente 500. Sin embargo, no se pretende que la presente invención se limite a un número específico de posturas consideradas.

VII. GENERACIÓN DE PROTEÍNAS CON ACTIVIDAD DESEADA MODIFICANDO SECUENCIAS PREDICADAS POR MODELO

[0224] Uno de los objetivos de la invención es generar una biblioteca de variantes de proteína optimizada a través de la evolución dirigida. Algunas realizaciones de la invención proporcionan métodos para guiar la evolución dirigida de variantes de proteínas usando los modelos de actividad de secuencia generados. Los diversos modelos de actividades de secuencia preparados y refinados de acuerdo con los métodos descritos anteriormente son adecuados para guiar la evolución dirigida de proteínas o moléculas biológicas. Como parte del proceso, los métodos pueden identificar secuencias que se utilizarán para generar nuevas variantes de proteínas para una próxima ronda de evolución dirigida como se indica en el bloque 111 de la Figura 1A. Dichas secuencias incluyen variaciones en los residuos definidos

identificados anteriormente, o son precursores utilizados para introducir posteriormente tales variaciones. Las secuencias pueden modificarse realizando mutagénesis y/o un mecanismo de generación de diversidad basado en recombinación para generar la nueva biblioteca de variantes de proteínas. En algunas realizaciones, las nuevas variantes pueden analizarse para determinar la actividad de interés. Ver el bloque 113 de la Figura 1A. En algunas aplicaciones, se pueden generar modelos estructurales para las nuevas variantes, que pueden proporcionar valores de energía y valores de geometría para las variantes. Ver el bloque 113 de la Figura 1A. En algunas realizaciones, estos datos pueden usarse luego para desarrollar un nuevo modelo de secuencia de actividad en una nueva ronda de evolución dirigida. Ver el bloque 115 de la Figura 1A.

[**0225]** En algunas realizaciones, la preparación de oligonucleótidos o secuencias de ácido nucleico se logra sintetizando los oligonucleótidos o secuencias de ácido nucleico usando un sintetizador de ácido nucleico. Algunas realizaciones de la invención incluyen realizar una ronda de evolución dirigida usando los oligonucleótidos preparados o la secuencia de proteínas como bloques de construcción para la evolución dirigida. Diversas realizaciones de la invención pueden aplicar recombinación y/o mutagénesis a estos bloques de construcción para generar diversidad.

[**0226]** En algunas realizaciones, el proceso identifica una o más secuencias que tienen propiedades ventajosas. A continuación, se generan variantes a partir de las secuencias identificadas como un conjunto de entrenamiento para un modelo de actividad de secuencia en una nueva ronda de evolución dirigida. Ver los bloques 355 y 357 de la figura 3C.

[**0227]** Para generar variantes, como un ejemplo específico, algunas realizaciones aplican técnicas de recombinación a oligonucleótidos. En estas realizaciones, los métodos implican seleccionar una o más mutaciones para una ronda de evolución dirigida evaluando los coeficientes de los términos del modelo de secuencia de actividad. Las mutaciones se seleccionan de combinaciones de aminoácidos o nucleótidos definidos de tipos de residuos específicos en posiciones específicas, en función de sus contribuciones a la actividad de las proteínas según lo predicho por los modelos. En algunas realizaciones, la selección de mutaciones implica identificar uno o más coeficientes que se determina que son mayores que otros de los coeficientes. Cada uno de los coeficientes se relaciona con la contribución de un residuo a la actividad proteica, y el residuo se define como de un tipo específico en una ubicación específica. La selección de mutaciones implica seleccionar los residuos asociados con uno o más coeficientes así identificados. En algunas realizaciones, después de seleccionar mutaciones de acuerdo con los modelos de secuencia de actividad, los métodos implican preparar una pluralidad de oligonucleótidos que contienen o codifican al menos una mutación, y realizar una ronda de evolución dirigida. En algunas realizaciones, las técnicas de evolución dirigida implican combinar y/o recombinar los oligonucleótidos.

[**0228]** Otras realizaciones aplican técnicas de recombinación a secuencias de proteínas. En algunas realizaciones, los métodos implican identificar una nueva proteína o una nueva secuencia de ácido nucleico, y preparar y analizar la nueva proteína o una proteína codificada por la nueva secuencia de ácido nucleico. En algunas realizaciones, los métodos implican además usar la nueva proteína o proteína codificada por la nueva secuencia de ácido nucleico como un punto de partida para una evolución dirigida adicional. En algunas realizaciones, el proceso de evolución dirigida implica fragmentar y recombinar la secuencia de proteína que el modelo predice que tiene un nivel de actividad deseado.

[**0229]** En algunas realizaciones, los métodos identifican y/o preparan una nueva proteína o una nueva secuencia de ácido nucleico en base a mutaciones individuales que el modelo predice que son importantes. Estos métodos implican: seleccionar una o más mutaciones mediante la evaluación de los coeficientes de los términos del modelo de secuencia de actividad para identificar uno o más de los aminoácidos o nucleótidos definidos en las posiciones definidas que contribuyen a la actividad; identificando una nueva proteína o una nueva secuencia de ácido nucleico que comprende la una o más mutaciones seleccionadas anteriormente, y preparando y analizando la nueva proteína o una proteína codificada por la nueva secuencia de ácido nucleico.

[**0230]** En otras realizaciones, los métodos identifican y/o preparan una nueva proteína o una nueva secuencia de ácido nucleico basándose en la actividad predicha de una secuencia completa en lugar de mutaciones individuales. En algunas de estas realizaciones, los métodos implican aplicar múltiples secuencias de proteínas o múltiples secuencias de aminoácidos al modelo de actividad de secuencia y determinar los valores de actividad predichos por el modelo de actividad de secuencia para cada una de las múltiples secuencias de proteína o secuencias de ácido nucleico. Los métodos implican además seleccionar una nueva secuencia de proteínas o una nueva secuencia de ácido nucleico de entre las múltiples secuencias de proteínas o múltiples secuencias de aminoácidos aplicadas anteriormente mediante la evaluación de los valores de actividad predichos por el modelo de secuencia de actividad para las múltiples secuencias. Los métodos también implican preparar y analizar la proteína que tiene la nueva secuencia de proteína o una proteína codificada por la nueva secuencia de ácido nucleico.

[**0231]** En algunas realizaciones, en lugar de simplemente sintetizar la proteína única mejor predicha, se genera una biblioteca combinatoria de proteínas basada en un análisis de sensibilidad de los mejores cambios en las elecciones de residuos en cada ubicación en la proteína. En esta realización, cuanto más sensible sea una elección de residuo dada para la proteína predicha, mayor será el cambio de aptitud previsto. En algunas realizaciones, estas sensibilidades son de mayor a menor y las puntuaciones de sensibilidad se usan para crear bibliotecas de proteínas

combinatorias en rondas posteriores (es decir, incorporando esos residuos en función de la sensibilidad). En algunas realizaciones, en las que se usa un modelo lineal/sin interacción, la sensibilidad se identifica simplemente considerando el tamaño del coeficiente asociado con un término de residuo dado en el modelo. Sin embargo, esto no es posible para modelos no lineales/de interacción. En cambio, en las realizaciones que utilizan modelos no lineales/de interacción, la sensibilidad del residuo se determina usando el modelo para calcular los cambios en la actividad cuando se varía un solo residuo en la "mejor" secuencia predicha.

[**0232]** Algunas realizaciones de la invención incluyen seleccionar una o más posiciones en la secuencia de proteínas o la secuencia de ácido nucleico y realizar mutagénesis de saturación en una o más posiciones así identificadas. En algunas realizaciones, las posiciones se seleccionan evaluando los coeficientes de los términos del modelo de secuencia de actividad para identificar uno o más de los aminoácidos o nucleótidos definidos en las posiciones definidas que contribuyen a la actividad. Por consiguiente, en algunas realizaciones, una ronda de evolución dirigida incluye realizar mutagénesis de saturación en una secuencia de proteína en posiciones seleccionadas usando los modelos de actividad de secuencia. En algunas realizaciones que implican modelos que comprenden uno o más términos de interacción, cada término de interacción se refiere a dos o más residuos. Los métodos implican aplicar mutagénesis simultáneamente en los dos o más residuos que interactúan.

[**0233]** En algunas realizaciones, los residuos se tienen en cuenta en el orden en que se clasifican. En algunas realizaciones, para cada residuo considerado, el proceso determina si se debe "alternar" ese residuo. El término "alternar" se refiere a incluir o excluir un residuo de aminoácido específico en una posición específica en las secuencias de variantes de proteínas en la biblioteca optimizada. Por ejemplo, la serina puede aparecer en la posición 166 en una variante de proteína, mientras que la fenilalanina puede aparecer en la posición 166 en otra variante de proteína en la misma biblioteca. Los residuos de aminoácidos que no varían entre las secuencias variantes de proteínas en el conjunto de entrenamiento generalmente permanecen fijos en la biblioteca optimizada. Sin embargo, este no es siempre el caso, ya que puede haber variación en las bibliotecas optimizadas.

[0234] En algunas realizaciones, una biblioteca de variantes de proteínas optimizada está diseñada de tal manera que todos los residuos de coeficiente de regresión de clasificación "alta" identificados son fijos, y el resto se alternan los residuos del coeficiente de regresión de menor rango. La justificación de esta realización es que se debe buscar el espacio local que rodea la "mejor" proteína predicha. Se observa que el punto de partida "columna vertebral" en donde se introducen los conmutadores puede ser la mejor proteína predicha por un modelo y/o una 'mejor' proteína ya validada de una biblioteca selectiva. De hecho, no se pretende que la columna vertebral del punto de partida se limite a ninguna proteína en particular.

[**0235]** En una realización alternativa, al menos uno o más, pero no todos los residuos de coeficiente de regresión de alto rango identificados se fijan en la biblioteca optimizada, y los demás se alternan. Este enfoque se recomienda en algunas realizaciones, si existe el deseo de no cambiar drásticamente el contexto de los otros residuos de aminoácidos incorporando demasiados cambios a la vez. Nuevamente, el punto de partida para alternar puede ser el mejor conjunto de residuos según lo predicho por el modelo, una proteína mejor validada de una biblioteca existente o un clon "promedio" que modela bien. En el último caso, puede ser deseable alternar los residuos que se prevé que sean de mayor importancia, ya que se debe explorar un espacio más grande en la búsqueda de colinas de actividad previamente omitidas del muestreo. Este tipo de biblioteca suele ser más relevante en las primeras rondas de producción de la biblioteca, ya que genera una imagen más refinada para las rondas posteriores. Tampoco se pretende que la columna vertebral del punto de partida se limite a ninguna proteína en particular.

[**0236]** Algunas alternativas de las realizaciones anteriores implican diferentes procedimientos para usar la importancia del residuo (es decir, clasificaciones) en la determinación de qué residuos alternar. En una de tales realizaciones alternativas, las posiciones de residuos de mayor clasificación se favorecen más agresivamente para alternar. La información necesaria en este enfoque incluye la secuencia de una mejor proteína del conjunto de entrenamiento, una mejor secuencia predicha por PLS o PCR, y una clasificación de los residuos del modelo PLS o PCR. En algunas realizaciones, la "mejor" proteína es un clon "mejor" validado en laboratorio húmedo en el conjunto de datos (es decir, el clon con la función medida más alta que todavía se modela bien porque cae relativamente cerca del valor predicho en la validación cruzada). El método compara cada residuo de esta proteína con el residuo correspondiente de una secuencia "mejor predicha" que tiene el valor más alto de la actividad deseada. Si el residuo con la carga más alta o el coeficiente de regresión no está presente en el 'mejor' clon, el método introduce esa posición como una posición de alternancia para la biblioteca posterior. Si el residuo está presente en el mejor clon, el método no trata la posición como una posición de alternancia, y se moverá a la siguiente posición sucesivamente. El proceso se repite para varios residuos, moviéndose a través de valores de carga sucesivamente más bajos, hasta que se genera una biblioteca de tamaño suficiente.

[**0237]** En algunas realizaciones adicionales, una proteína 'mejor' (o una de las mejores) validada en laboratorio húmedo en la biblioteca optimizada actual (es decir, una proteína con la función medida más alta o una de las más altas que todavía se modela bien), es decir, cae relativamente cerca del valor predicho en la validación cruzada) sirve como columna vertebral en la que se incorporan varios cambios. En otro enfoque, una proteína 'mejor' (o una de las mejores) validada en laboratorio húmedo en la biblioteca actual que puede no modelar bien sirve como una columna vertebral donde se incorporan varios cambios. En algunos otros enfoques, una secuencia predicha por el modelo de

actividad de secuencia para tener el valor más alto (o uno de los valores más altos) de la actividad deseada sirve como la columna vertebral. En estos enfoques, el conjunto de datos para la biblioteca de "próxima generación" (y posiblemente un modelo correspondiente) se obtiene cambiando los residuos en al menos una de las mejores proteínas. En una realización, estos cambios comprenden una variación sistemática de los residuos en la cadena principal. En algunos casos, los cambios comprenden diversas técnicas de mutagénesis, recombinación y/o selección de subsecuencias. Cada uno de estos se puede realizar *in vitro, in vivo* y/o *in silico.* De hecho, no se pretende que la presente divulgación se limite a ningún formato particular, ya que cualquier formato adecuado encuentra uso.

[**0238]** En algunas realizaciones, las bibliotecas de variantes de proteínas optimizadas se generan usando los métodos de recombinación descritos en este documento, o alternativamente, mediante métodos de síntesis de genes, seguidos de expresión *in vivo* o *in vitro.* En algunas realizaciones, después de que las bibliotecas de variantes de proteínas optimizadas se seleccionan para la actividad deseada, se secuencian. Como se indicó anteriormente, la información de actividad y secuencia de la biblioteca de variantes de proteína optimizada se puede emplear para generar otro modelo de actividad de secuencia a partir del cual se puede diseñar una biblioteca optimizada adicional, usando los métodos descritos en este documento. En una realización, todas las proteínas de esta nueva biblioteca se usan como parte del conjunto de datos.

VIII SECUENCIACIÓN DE POLINUCLEÓTIDOS Y POLIPÉPTIDOS

[0239] En algunas realizaciones, la información de secuencia de polinucleótidos y polipéptidos se usa para generar modelos de actividad de secuencia o representaciones computacionales de sitios activos de variantes de proteínas. En algunas realizaciones, la información de secuencia de polinucleótidos y polipéptidos se usa en procesos de evolución dirigida para obtener variantes de proteínas de propiedades deseadas.

[**0240]** En diversas realizaciones, las secuencias de variantes de proteínas se determinan a partir de biomoléculas físicas mediante métodos de secuenciación de proteínas, algunos de los cuales se describen adicionalmente a continuación. La secuenciación de proteínas implica determinar la secuencia de aminoácidos de una proteína. Algunas técnicas de secuenciación de proteínas también determinan la conformación que adopta la proteína y el grado en que se compleja con cualquier molécula no peptídica. La espectrometría de masas y la reacción de degradación de Edman pueden usarse para determinar directamente la secuencia de aminoácidos de una proteína.

[**0241]** La reacción de degradación de Edman permite descubrir la composición de aminoácidos ordenada de una proteína. En algunos ejemplos, se pueden usar secuenciadores Edman automatizados para determinar la secuencia de variantes de proteínas. Los secuenciadores Edman automatizados pueden secuenciar péptidos de secuencias cada vez más largas, por ejemplo, de hasta aproximadamente 50 aminoácidos de longitud. En algunas realizaciones, un proceso de secuenciación de proteínas que implementa la degradación de Edman implica uno o más de los siguientes:

- Romper puentes de disulfuro en la proteína con un agente reductor, por ejemplo, 2-mercaptoetanol. Se puede usar un grupo protector como el ácido yodoacético para evitar que se vuelvan a formar los enlaces.
- Separar y purificar las cadenas individuales del complejo proteico si hay más de una.
- Determinar la composición de aminoácidos de cada cadena
- Determinar los aminoácidos terminales de cada cadena
- Romper cada cadena en fragmentos, por ejemplo, fragmentos de menos de 50 aminoácidos de largo.
- Separar y purificar los fragmentos
- Determinar la secuencia de cada fragmento utilizando la reacción de degradación de Edman
- Repetir los pasos anteriores aplicando un patrón diferente de escisión para proporcionar lecturas adicionales de secuencias de aminoácidos
- Construir la secuencia de proteína global de lecturas de la secuencia de aminoácidos.

[**0242]** En diversas implementaciones, los péptidos más largos que aproximadamente 50-70 aminoácidos deben dividirse en pequeños fragmentos para facilitar la secuenciación por las reacciones de Edman. La digestión de secuencias más largas puede realizarse mediante endopeptidasas como la tripsina o la pepsina, o mediante reactivos químicos como el bromuro de cianógeno. Diferentes enzimas dan diferentes patrones de escisión, y la superposición entre fragmentos se puede usar para construir una secuencia general.

[**0243]** Durante la reacción de degradación de Edman, el péptido a secuenciar se adsorbe sobre una superficie sólida de un sustrato. En algunos ejemplos, un sustrato adecuado es la fibra de vidrio recubierta con polibreno, un polímero catiónico. El reactivo de Edman, fenilisotiocianato (PITC), se agrega al péptido adsorbido, junto con una solución tampón de trimetilamina ligeramente básica. Esta solución de reacción reacciona con el grupo amina del aminoácido N-terminal. El aminoácido terminal se puede desprender selectivamente mediante la adición de ácido anhidro. El derivado luego se isomeriza para dar una feniltiohidantoína sustituida, que se puede lavar e identificar por cromatografía. Entonces el ciclo puede repetirse.

[**0244]** En algunos ejemplos, la espectrometría de masas se puede usar para determinar una secuencia de aminoácidos determinando las relaciones masa-carga de fragmentos de la secuencia de aminoácidos. Se puede

determinar el espectro de masas que incluye picos correspondientes a fragmentos con carga múltiple, donde la distancia entre los picos correspondientes a diferentes isótopos es inversamente proporcional a la carga en el fragmento. El espectro de masas se analiza, por ejemplo, mediante comparación con una base de datos de proteínas secuenciadas previamente para determinar las secuencias de los fragmentos. Este proceso se repite con una enzima de digestión diferente, y las superposiciones en las secuencias se utilizan para construir una secuencia de aminoácidos completa.

[**0245]** Los péptidos son a menudo más fáciles de preparar y analizar para la espectrometría de masas de proteínas enteras. En algunos ejemplos, la ionización por electropulverización se usa para administrar los péptidos al espectrómetro. La proteína es digerida por una endoproteasa, y la solución resultante se pasa a través de una columna de cromatografía líquida de alta presión. Al final de esta columna, la solución se rocía en el espectrómetro de masas, la solución se carga con un potencial positivo. La carga en las gotas de solución hace que se fragmenten en iones individuales. Los péptidos se fragmentan y se miden las relaciones de masa a carga de los fragmentos.

[**0246]** También es posible determinar indirectamente una secuencia de aminoácidos a partir de la secuencia de ADN o ARNm que codifica la proteína. Los métodos de secuenciación de ácido nucleico, por ejemplo, varios métodos de secuenciación de próxima generación, pueden usarse para determinar secuencias de ADN o ARN. En algunas implementaciones, una secuencia de proteína se aísla nuevamente sin conocimiento de los nucleótidos que codifican la proteína. En tales implementaciones, primero se puede determinar una secuencia corta de polipéptidos usando uno de los métodos de secuenciación directa de proteínas. Se puede determinar un marcador complementario para el ARN de la proteína a partir de esta secuencia corta. Esto se puede usar para aislar el ARNm que codifica la proteína, que luego se puede replicar en una reacción en cadena de la polimerasa para producir una cantidad significativa de ADN, que luego se puede secuenciar usando métodos de secuenciación de ADN. La secuencia de aminoácidos de la proteína se puede deducir de la secuencia de ADN. En la deducción, es necesario tener en cuenta los aminoácidos eliminados después de que se haya traducido el ARNm.

[**0247]** En diversas realizaciones, la información de secuencia de polinucleótidos se usa para generar modelos de actividad de secuencia o representación computacional de sitios de actividad de proteína. La información de la secuencia de ácido nucleico se puede determinar a partir de biomoléculas físicas mediante métodos de secuenciación de ácido nucleico, algunos de los cuales se describen adicionalmente a continuación.

[**0248]** En una o más realizaciones, los datos de secuencia se pueden obtener usando métodos de secuenciación masiva que incluyen, por ejemplo, secuenciación Sanger o secuenciación Maxam-Gilbert, que se consideran los métodos de secuenciación de primera generación. La secuenciación de Sanger, que implica el uso de terminadores de cadena didesoxi marcados, es bien conocida en la técnica; véase, por ejemplo, Sanger et al., Proceedings of the National Academy of Sciences of the United States of America 74, 5463-5467 (1997). La secuenciación de Maxam-Gilbert, que implica realizar múltiples reacciones de degradación química parcial en fracciones de la muestra de ácido nucleico seguida de detección y análisis de los fragmentos para inferir la secuencia, también es conocida en la técnica; véase, por ejemplo, Maxam et al., Proceedings of the National Academy of Sciences of the United States of America 74, 560-564 (1977). Otro método de secuenciación masiva es la secuenciación por hibridación, en la que la secuencia de una muestra se deduce en función de sus propiedades de hibridación a una pluralidad de secuencias, por ejemplo, en un microarray o chip genético; véase, por ejemplo, Drmanac, et al., Nature Biotechnology 16, 54-58 (1998).

[**0249]** En una o más realizaciones, los datos de secuencia se obtienen usando métodos de secuenciación de próxima generación. La secuenciación de próxima generación también se conoce como secuenciación de alto rendimiento. Las técnicas paralelizan el proceso de secuenciación, produciendo miles o millones de secuencias a la vez. Los ejemplos de métodos de secuenciación de próxima generación adecuados incluyen, entre otros, secuenciación en tiempo real de una sola molécula (p. ej., Pacific Biosciences de Menlo Park, California), secuenciación de semiconductores iónicos (p. ej., Ion Torrent de South San Francisco, California), secuenciación de pirosis (p. ej., 454 de Branford, Connecticut), secuenciación por ligadura (p. ej., secuenciación SOLiD propiedad de Life Technologies de Carlsbad, California), secuenciación por síntesis y terminador reversible (p. ej., Illumina de San Diego, California), tecnologías de imágenes de ácidos nucleicos tales como microscopía electrónica de transmisión y similares.

[**0250]** En general, los métodos de secuenciación de próxima generación típicamente usan un paso de clonación *in vitro* para amplificar moléculas de ADN individuales. La PCR de emulsión (emPCR) aísla moléculas de ADN individuales junto con perlas recubiertas con cebador en gotas acuosas dentro de una fase oleosa. La PCR produce copias de la molécula de ADN, que se une a los cebadores en la cuenta, seguido de la inmovilización para una secuenciación posterior. emPCR se utiliza en los métodos de Marguilis *et al.* (comercializado por 454 Life Sciences, Branford, CT), Shendure y Porreca *et al.* (también conocido como "secuenciación de polonia") y secuenciación SOLiD (Applied Biosystems Inc., Foster City, CA). Ver M. Margulies, et al. (2005) "Genome sequencing in microfabricated high-density picolitre reactors" Nature 437: 376-380; J. Shendure y col. (2005) "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome" Science 309 (5741): 1728-1732. La amplificación clonal in vitro también se puede llevar a cabo mediante "PCR puente", donde los fragmentos se amplifican sobre cebadores unidos a una superficie sólida. Braslavsky y col. desarrolló un método de molécula única (comercializado por Helicos Biosciences Corp., Cambridge, MA) que omite este paso de amplificación, fijando directamente las moléculas de ADN a una superficie. I.

Braslavsky y col. (2003) "Sequence information can be obtained from single DNA molecules" Proceedings of the National Academy of Sciences of the United States of America 100: 3960-3964.

[**0251]** Las moléculas de ADN que se unen físicamente a una superficie pueden secuenciarse en paralelo. En la "secuenciación por síntesis", se construye una cadena complementaria basada en la secuencia de una cadena de plantilla que usa una ADN polimerasa. como la secuenciación electroforética con terminación de tinte, los métodos de terminación reversible (comercializados por Illumina, Inc., San Diego, CA y Helicos Biosciences Corp., Cambridge, MA) utilizan versiones reversibles de terminadores de tintura, agregando un nucleótido a la vez, y detectan fluorescencia en cada posición en tiempo real, mediante la eliminación repetida del grupo de bloqueo para permitir la polimerización de otro nucleótido. La "pirosecuenciación" también utiliza la polimerización de ADN, agregando un nucleótido a la vez y detectando y cuantificando el número de nucleótidos agregados a una ubicación dada a través de la luz emitida por la liberación de pirofosfatos unidos (comercializado por 454 Life Sciences, Branford, CT). Ver M. Ronaghi, et al. (1996) "Real-time DNA sequencing using detection of pyrophosphate release" Analytical Biochemistry 242: 84-89.

[**0252]** Ejemplos específicos de los métodos de secuenciación de próxima generación se describen con más detalles a continuación. Una o más implementaciones de la presente invención pueden usar uno o más de los siguientes métodos de secuenciación sin desviarse de los principios de la invención.

[**0253]** La secuenciación en tiempo real de una sola molécula (también conocida como SMRT) es una secuenciación de ADN de una sola molécula paralela por tecnología de síntesis desarrollada por Pacific Biosciences. La secuenciación en tiempo real de una sola molécula utiliza la guía de onda de modo cero (ZMW). Una única enzima ADN polimerasa se fija en la parte inferior de un ZMW con una sola molécula de ADN como plantilla. El ZMW es una estructura que crea un volumen de observación iluminado que es lo suficientemente pequeño como para observar un solo nucleótido de ADN (también conocido como base) que se incorpora por la ADN polimerasa. Cada una de las cuatro bases de ADN está unida a uno de los cuatro tintes fluorescentes diferentes. Cuando la ADN polimerasa incorpora un nucleótido, la etiqueta fluorescente se separa y se difunde fuera del área de observación del ZMW donde su fluorescencia ya no es observable. Un detector detecta la señal fluorescente de la incorporación de nucleótidos, y la llamada de base se realiza de acuerdo con la fluorescencia correspondiente del colorante.

[**0254]** Otra tecnología de secuenciación de molécula única aplicable es la tecnología Helicos True Single Molecule Sequencing (tSMS) (por ejemplo, como se describe en Harris TD et al., Science 320: 106-109 [2008]). En la técnica tSMS, una muestra de ADN se divide en cadenas de aproximadamente 100 a 200 nucleótidos, y se agrega una secuencia de poliA al extremo 3' de cada cadena de ADN. Cada cadena se marca mediante la adición de un nucleótido de adenosina marcado con fluorescencia. Las cadenas de ADN luego se hibridan con una célula de flujo, que contiene millones de sitios de captura de oligo-T que se inmovilizan en la superficie de la célula de flujo. En ciertas realizaciones, las plantillas pueden tener una densidad de aproximadamente 100 millones de plantillas/cm^2. La célula de flujo se carga en un instrumento, por ejemplo, el secuenciador HeliScope™, y un láser ilumina la superficie de la célula de flujo, revelando la posición de cada plantilla. Una cámara CCD puede asignar la posición de las plantillas en la superficie de la célula de flujo. La plantilla de etiqueta fluorescente se escinde y se lava. La reacción de secuenciación comienza introduciendo una ADN polimerasa y un nucleótido marcado con fluorescencia. El ácido nucleico oligo-T sirve como cebador. La polimerasa incorpora los nucleótidos marcados al cebador de una manera dirigida por plantilla. La polimerasa y los nucleótidos no incorporados se eliminan. Las plantillas que han dirigido la incorporación del nucleótido marcado con fluorescencia se distinguen mediante la formación de imágenes de la superficie de la célula de flujo. Después de la formación de imágenes, una etapa de escisión elimina la etiqueta fluorescente, y el proceso se repite con otros nucleótidos marcados con fluorescencia hasta que se alcanza la longitud de lectura deseada. La información de secuencia se recopila con cada paso de adición de nucleótidos. La secuenciación del genoma completo mediante tecnologías de secuenciación de una sola molécula excluye u obvia la amplificación basada en PCR en la preparación de las bibliotecas de secuenciación, y los métodos permiten la medición directa de la muestra, en lugar de la medición de copias de esa muestra.

[**0255]** La secuenciación de semiconductores de iones es un método de secuenciación de ADN basado en la detección de iones de hidrógeno que se liberan durante la polimerización de ADN. Este es un método de "secuenciación por síntesis", durante el cual se construye una cadena complementaria basada en la secuencia de una cadena de plantilla. Un micropocillo que contiene una cadena de ADN plantilla para ser secuenciado se inunda con una sola especie de desoxirribonucleótido trifosfato (dNTP). Si el dNTP introducido es complementario al nucleótido molde principal, se incorpora a la cadena complementaria en crecimiento. Esto provoca la liberación de un ion de hidrógeno que activa un sensor de iones ISFET, lo que indica que se ha producido una reacción. Si las repeticiones de homopolímero están presentes en la secuencia de plantilla, se incorporarán múltiples moléculas de dNTP en un solo ciclo. Esto conduce a un número correspondiente de hidrógenos liberados y una señal electrónica proporcionalmente más alta. Esta tecnología difiere de otras tecnologías de secuenciación en que no se utilizan nucleótidos u ópticos modificados. La secuenciación de semiconductores iónicos también puede denominarse secuenciación de torrente iónico, secuenciación mediada por pH, secuenciación de silicio o secuenciación de semiconductores.

[**0256]** En la pirosecuenciación, el ion pirofosfato liberado por la reacción de polimerización se hace reaccionar con adenosina 5' fosfosulfato por ATP sulfurilasa para producir ATP; el ATP impulsa la conversión de luciferina a oxilciferina

más luz por luciferasa. Como la fluorescencia es transitoria, no es necesario un paso separado para eliminar la fluorescencia en este método. Se agrega un tipo de desoxirribonucleótido trifosfato (dNTP) a la vez, y la información de la secuencia se distingue según la cual dNTP genera una señal significativa en un sitio de reacción. El instrumento Roche GS FLX disponible en el mercado adquiere la secuencia utilizando este método. Esta técnica y sus aplicaciones se analizan en detalle, por ejemplo, en Ronaghi et al., Analytical Biochemistry 242, 84-89 (1996) y Margulies et al., Nature 437, 376-380 (2005) (corrigendum en Nature 441, 120 (2006)). Una tecnología de pirosecuenciación disponible comercialmente es la secuenciación 454 (Roche) (por ejemplo, como se describe en Margulies, M. et al. Nature 437: 376-380 [2005]).

[**0257]** En la secuenciación de la ligadura, se usa una enzima ligasa para unir un oligonucleótido parcialmente bicatenario con un saliente al ácido nucleico que se está secuenciando, que tiene un saliente; Para que se produzca la ligadura, los voladizos deben ser complementarios. Las bases en el saliente del oligonucleótido parcialmente bicatenario se pueden identificar de acuerdo con un fluoróforo conjugado con el oligonucleótido parcialmente bicatenario y/o un oligonucleótido secundario que hibrida con otra parte del oligonucleótido parcialmente bicatenario. Después de la adquisición de los datos de fluorescencia, el complejo ligado se escinde aguas arriba del sitio de ligadura, como por ejemplo una enzima de restricción de tipo II, por ejemplo, Bbvl, que corta en un sitio una distancia fija de su sitio de reconocimiento (que se incluyó parcialmente oligonucleótido bicatenario). Esta reacción de escisión expone un nuevo voladizo aguas arriba del voladizo anterior, y el proceso se repite. Esta técnica y sus aplicaciones se discuten en detalle, por ejemplo, en Brenner et al., Nature Biotechnology 18, 630-634 (2000). En algunas realizaciones, la secuenciación de ligadura se adapta a los métodos de la invención obteniendo un producto de amplificación de círculo rodante de una molécula de ácido nucleico circular, y usando el producto de amplificación de círculo rodante como plantilla para la secuenciación de ligadura.

[**0258]** Un ejemplo comercialmente disponible de tecnología de secuenciación de ligadura es la tecnología SOLiD™ (Applied Biosystems). En la secuenciación SOLiD™ por ligadura, el ADN genómico se corta en fragmentos, y los adaptadores se unen a los extremos 5' y 3' de los fragmentos para generar una biblioteca de fragmentos. Alternativamente, los adaptadores internos se pueden introducir ligando los adaptadores a los extremos 5' y 3' de los fragmentos, circularizando los fragmentos, digiriendo el fragmento circularizado para generar un adaptador interno, y uniendo los adaptadores a los extremos 5' y 3' del resultado fragmentos para generar una biblioteca emparejada por parejas. A continuación, las poblaciones de microesferas clonales se preparan en microrreactores que contienen microesferas, cebadores, plantilla y componentes de PCR. Después de la PCR,las plantillas se desnaturalizan y las cuentas se enriquecen para separar las cuentas con plantillas extendidas. Las plantillas en las cuentas seleccionadas se someten a una modificación de 3' que permite la unión a un portaobjetos de vidrio. La secuencia se puede determinar mediante hibridación secuencial y ligadura de oligonucleótidos parcialmente aleatorios con una base determinada central (o un par de bases) que se identifica por un fluoróforo específico. Después de registrar un color, el oligonucleótido ligado se escinde y se elimina y luego se repite el proceso.

[**0259]** En la secuenciación del terminador reversible, se incorpora un análogo de nucleótido marcado con colorante fluorescente que es un terminador de cadena reversible debido a la presencia de un grupo bloqueante en una reacción de extensión de base única. La identidad de la base se determina según el fluoróforo; en otras palabras, cada base se combina con un fluoróforo diferente. Después de adquirir los datos de secuencia/fluorescencia, el fluoróforo y el grupo de bloqueo se eliminan químicamente, y el ciclo se repite para adquirir la siguiente base de información de secuencia. El instrumento Illumina GA funciona por este método. Esta técnica y sus aplicaciones se analizan en detalle, por ejemplo, en Ruparel et al., Proceedings of the National Academy of Sciences of the United States of America 102, 5932-5937 (2005), y Harris et al., Science 320, 106-109 (2008).

[**0260]** Un ejemplo comercialmente disponible del método de secuenciación del terminador reversible es la secuenciación por síntesis de Illumina y la secuenciación basada en el terminador reversible (por ejemplo, como se describe en Bentley et al., Nature 6: 53-59 [2009]). La tecnología de secuenciación de Illumina se basa en la unión de ADN genómico fragmentado a una superficie plana, ópticamente transparente, sobre la cual se unen los anclajes de oligonucleótidos. El ADN de plantilla se repara en el extremo para generar extremos romos fosforilados en 5', y la actividad de polimerasa del fragmento Klenow se usa para agregar una sola base A al extremo 3' de los fragmentos de ADN fosforilados romos. Esta adición prepara los fragmentos de ADN para la unión a adaptadores de oligonucleótidos, que tienen un saliente de una única base T en su extremo 3' para aumentar la eficiencia de la ligadura. Los oligonucleótidos adaptadores son complementarios a los anclajes de las células de flujo. En condiciones de dilución limitante, se agrega ADN de plantilla monocatenario modificado por adaptador a la célula de flujo y se inmoviliza por hibridación a los anclajes. Los fragmentos de ADN unidos se extienden y amplifican en puente para crear una célula de flujo de secuenciación de densidad ultraalta con cientos de millones de grupos, cada uno con ~ 1.000 copias de la misma plantilla. Las plantillas se secuencian utilizando una tecnología robusta de secuenciación de ADN de cuatro colores por síntesis que emplea terminadores reversibles con tintes fluorescentes extraíbles. La detección de fluorescencia de alta sensibilidad se logra utilizando excitación láser y óptica de reflexión interna total. Las lecturas de secuencia corta de aproximadamente 20-40 pb, por ejemplo, 36 pb, se alinean contra un genoma de referencia enmascarado repetidamente y el mapeo único de las lecturas de secuencia corta al genoma de referencia se identifica usando un software de canalización de análisis de datos especialmente desarrollado. También se pueden usar genomas de referencia enmascarados sin repetición. Ya sea que se utilicen genomas de referencia enmascarados repetidos o no enmascarados, solo se cuentan las lecturas que se asignan de forma exclusiva al

genoma de referencia. Después de completar la primera lectura, las plantillas se pueden regenerar *in situ* para permitir una segunda lectura desde el extremo opuesto de los fragmentos. Por lo tanto, se puede usar la secuenciación de extremo único o de extremo emparejado de los fragmentos de ADN. Se realiza la secuenciación parcial de fragmentos de ADN presentes en la muestra, y se cuentan las etiquetas de secuencia que comprenden lecturas de longitud predeterminada, por ejemplo, 36 pb, a un genoma de referencia conocido.

[**0261]** En la secuenciación de nanoporos, una molécula de ácido nucleico monocatenario se pasa a través de un poro, por ejemplo, usando una fuerza impulsora electroforética, y la secuencia se deduce analizando los datos obtenidos a medida que la molécula de ácido nucleico monocatenario pasa a través del poro. Los datos pueden ser datos de corriente iónica, en donde cada base altera la corriente, por ejemplo, bloqueando parcialmente la corriente que pasa a través del poro en un grado diferente y distinguible.

[**0262]** En otra realización ilustrativa, pero no limitante, los métodos descritos en el presente documento comprenden obtener información de secuencia usando microscopía electrónica de transmisión (TEM). El método comprende utilizar imágenes de microscopio electrónico de transmisión de resolución de un solo átomo de ADN de alto peso molecular (150 kb o más) marcado selectivamente con marcadores de átomos pesados y organizar estas moléculas en películas ultrafinas en matrices paralelos ultradensos (cadena a cadena de 3 nm) con espaciado constante de base a base. El microscopio electrónico se usa para obtener imágenes de las moléculas en las películas para determinar la posición de los marcadores de átomos pesados y para extraer información de la secuencia de bases del ADN. El método se describe adicionalmente en la publicación de patente PCT WO 2009/046445.

[**0263]** En otra realización ilustrativa, pero no limitativa, los métodos descritos en el presente documento comprenden obtener información de secuencia usando secuenciación de tercera generación. En la secuenciación de tercera generación, se usa un portaobjetos con un revestimiento de aluminio con muchos agujeros pequeños (~ 50 nm) como guía de onda en modo cero (véase, por ejemplo, Levene et al., Science 299, 682-686 (2003)). La superficie de aluminio está protegida de la unión de la ADN polimerasa por la química del polifosfonato, por ejemplo, la química del polivinilfosfonato (véase, por ejemplo, Korlach et al., Proceedings of the National Academy of Sciences of the United States of America 105, 1176-1181 (2008)) Esto da como resultado una unión preferencial de las moléculas de ADN polimerasa a la sílice expuesta en los agujeros del revestimiento de aluminio. Esta configuración permite que se usen fenómenos de ondas evanescentes para reducir el fondo de fluorescencia, lo que permite el uso de concentraciones más altas de dNTP marcados con fluorescencia. El fluoróforo se une al fosfato terminal de los dNTP, de modo que se libera fluorescencia al incorporarse el dNTP, pero el fluoróforo no permanece unido al nucleótido recién incorporado, lo que significa que el complejo está inmediatamente listo para otra ronda de incorporación. Mediante este método, se puede detectar la incorporación de dNTP en complejos de plantilla de cebador individuales presentes en los agujeros del revestimiento de aluminio. Ver, por ejemplo, Eid et al., Science 323, 133-138 (2009).

IX. ENSAYO DE VARIANTES DE GENES Y PROTEÍNAS

[0264] En algunas realizaciones, los polinucleótidos generados en conexión con los métodos de la presente invención se clonan opcionalmente en células para expresar variantes de proteínas para el cribado de actividad (o se usan en reacciones de transcripción *in vitro* para fabricar productos que se criban). Además, los ácidos nucleicos que codifican las variantes de proteínas se pueden enriquecer, secuenciar, expresar, amplificar *in vitro* o tratar en cualquier otro método recombinante común.

[**0265]** Los textos generales que describen técnicas de biología molecular útiles en el presente documento, que incluyen clonación, mutagénesis, construcción de bibliotecas, ensayos de detección, cultivo celular y similares, incluyen Berger y Kimmel, Guía de técnicas de clonación molecular, Métodos en enzimología, 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 Current Protocols in Molecular Biology, FM 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)). Los métodos de transducción de células, incluidas las células vegetales y animales, con ácidos nucleicos están generalmente disponibles, al igual que los métodos de expresión de proteínas codificadas por dichos ácidos nucleicos. Además de Berger, Ausubel y Sambrook, las referencias generales útiles para el cultivo de células animales incluyen Freshney (Culture of Animal Cells, Manual of Basic Technique, tercera edición Wiley-Liss, Nueva York (1994)) y las referencias allí citadas, Humason (Animal Tissue Techniques, cuarta edición WH Freeman and Company (1979)) y Ricciardelli, et al., In Vitro Cell Dev. Biol. 25: 1016 - 1024 (1989). Las referencias para la clonación, cultivo y regeneración de células vegetales incluyen Payne et al. (1992) Cultivo de células vegetales y tejidos en sistemas líquidos John Wiley & Sons, Inc. Nueva York, NY (Payne); y Gamborg y Phillips (eds) (1995) Cultivo de células vegetales, tejidos y órganos; Métodos fundamentales Springer Lab Manual, Springer-Verlag (Berlín Heidelberg Nueva York) (Gamborg). Una variedad de medios de cultivo celular se describe en Atlas y Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL (Atlas). Se encuentra información adicional para el cultivo de células vegetales en la literatura comercial disponible, tal como el Life Science Research Cell Culture Catalog (1998) de Sigma-Aldrich, Inc. (St Louis, MO) (Sigma-LSRCCC) y, por ejemplo, el Plant Culture Catalog y suplemento (1997) también de SigmaAldrich, Inc (St Louis, MO) (Sigma-PCCS).

[**0266]** Ejemplos de técnicas suficientes para dirigir a personas expertas a través de métodos de amplificación *in vitro,*

útiles, por ejemplo, para amplificar ácidos nucleicos recombinados con oligonucleótidos que incluyen reacciones en cadena de la polimerasa (PCR), reacciones en cadena de la ligasa (LCR), amplificaciones de Qβ-replicasa y otras técnicas mediadas por ARN polimerasa (p. ej., 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. San Diego, CA (1990) (Innis); Arnheim y Levinson (1 de octubre de 1990) C&EN 36-47; The Journal Of NIH Research (1991) 3, 81-94; Kwoh y col. (1989) Proc. Natl. Acad. Sci. Estados Unidos 86, 1173; Guatelli y col. (1990) Proc. Natl. Acad. Sci. Estados Unidos 87, 1874; Lomell y col. (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) Gene 4, 560; Barringer y col. (1990) Gene 89, 117 y Sooknanan y Malek (1995) Biotechnology 13: 563-564. Los métodos mejorados de clonación de ácidos nucleicos amplificados *in vitro* se describen en Wallace et al., Patente de EE.UU. Nº 5.426.039. Los métodos mejorados para amplificar ácidos nucleicos grandes por PCR se resumen en Cheng et al. (1994) Nature 369: 684-685 y sus referencias, en las que se generan amplicones de PCR de hasta 40 kb. Un experto apreciará que, esencialmente, cualquier ARN puede convertirse en un ADN bicatenario adecuado para la digestión de restricción, la expansión de PCR y la secuenciación utilizando transcriptasa inversa y una polimerasa. Ver, Ausubel, Sambrook y Berger, *todos supra.*

[0267] En un método preferido, las secuencias reensambladas se verifican para la incorporación de oligonucleótidos de recombinación basados en la familia. Esto puede hacerse clonando y secuenciando los ácidos nucleicos, y/o mediante digestión de restricción, por ejemplo, como se enseña esencialmente en Sambrook, Berger y Ausubel, *supra.* Además, las secuencias pueden amplificarse por PCR y secuenciarse directamente. Por lo tanto, además de, por ejemplo, Sambrook, Berger, Ausubel e Innis (*supra)*, las metodologías de secuenciación de PCR adicionales también son particularmente útiles. Por ejemplo, la secuenciación directa de amplicones generados por PCR incorporando selectivamente nucleótidos resistentes a nucleasas boronadas en los amplicones durante la PCR y la digestión de los amplicones con una nucleasa para producir fragmentos de plantilla dimensionados (Porter et al. (1997) Nucleic Acids Research 25 (8): 1611-1617). En los métodos, se realizan cuatro reacciones de PCR en una plantilla, en cada una de las cuales uno de los nucleótidos trifosfatos en la mezcla de reacción de PCR está parcialmente sustituido con un 2'desoxinucleósido 5'-[P-borano]-trifosfato. El nucleótido boronado se incorpora estocásticamente a los productos de PCR en diferentes posiciones a lo largo del amplicón de PCR en un conjunto anidado de fragmentos de PCR de la plantilla. Una exonucleasa que está bloqueada por nucleótidos boronados incorporados se usa para escindir los amplicones de PCR. Los amplicones escindidos se separan por tamaño utilizando electroforesis en gel de poliacrilamida, proporcionando la secuencia del amplicón. Una ventaja de este método es que utiliza menos manipulaciones bioquímicas que la secuenciación estándar de Sanger de amplicones de PCR.

[0268] Los genes sintéticos son susceptibles de clonación convencional y enfoques de expresión; así, las propiedades de los genes y proteínas que codifican pueden examinarse fácilmente después de su expresión en una célula huésped. Los genes sintéticos también se pueden usar para generar productos de polipéptidos mediante transcripción y traducción *in vitro* (sin células). Por lo tanto, los polinucleótidos y polipéptidos pueden examinarse para determinar su capacidad para unir una variedad de ligandos predeterminados, moléculas e iones pequeños, o sustancias poliméricas y heteropoliméricas, incluidas otras proteínas y epítopos de polipéptidos, así como paredes celulares microbianas, partículas virales, superficies y membranas.

[0269] Por ejemplo, se pueden usar muchos métodos físicos para detectar polinucleótidos que codifican fenotipos asociados con la catálisis de reacciones químicas por polinucleótidos directamente o por polipéptidos codificados. Con el único fin de ilustrar, y dependiendo de los detalles de reacciones químicas predeterminadas particulares de interés, estos métodos pueden incluir una multitud de técnicas conocidas en la técnica que explican una diferencia física entre sustrato(s) y producto(s), o para cambios en los medios de reacción asociados con la reacción química (por ejemplo, cambios en las emisiones electromagnéticas, adsorción, disipación y 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 isotópicamente etiquetados, partición y métodos espectrales que representan la distribución de isótopos o la formación de productos etiquetados ; métodos espectrales y químicos para detectar cambios acompañantes en iones o composiciones elementales de producto(s) de reacción (incluyendo cambios en pH, iones inorgánicos y orgánicos y similares). Otros métodos de ensayos físicos, adecuados para su uso en los métodos de la presente memoria, pueden basarse en el uso de biosensores específicos para los productos de reacción, incluidos los que comprenden anticuerpos con propiedades indicadoras, o los basados en el reconocimiento de afinidad *in vivo* junto con la expresión y actividad de un gen reportero. Los ensayos acoplados a enzimas para la detección del producto de reacción y las selecciones de crecimiento de vida-muerte celular *in vivo* también se pueden usar cuando sea apropiado. Independientemente de la naturaleza específica de los ensayos físicos, todos se usan para seleccionar una actividad deseada, o combinación de actividades deseadas, proporcionadas o codificadas por una biomolécula de interés.

[0270] El ensayo específico utilizado para la selección dependerá de la aplicación. Se conocen muchos ensayos para proteínas, receptores, ligandos, enzimas, sustratos y similares. Los formatos incluyen la unión a componentes inmovilizados, la viabilidad celular u orgánica, la producción de composiciones indicadoras y similares.

[0271] Ensayos de alto rendimiento son particularmente adecuados para el cribado de bibliotecas empleadas en la presente invención. En ensayos de alto rendimiento, es posible detectar hasta varios miles de variantes diferentes en

un solo día. Por ejemplo, cada pocillo de una placa de microtitulación se puede usar para realizar un ensayo por separado o, si se observan efectos de concentración o tiempo de incubación, cada 5-10 pocillos pueden probar una única variante (p. ej., a diferentes concentraciones). Por lo tanto, una única placa de microtitulación estándar puede analizar aproximadamente 100 (por ejemplo, 96) reacciones. Si se usan placas de 1536 pocillos, entonces una sola placa puede analizar fácilmente entre aproximadamente 100 y aproximadamente 1500 reacciones diferentes. Es posible analizar varias placas diferentes por día; los cribados de ensayo para hasta aproximadamente 6.000-20.000 ensayos diferentes (es decir, que implican diferentes ácidos nucleicos, proteínas codificadas, concentraciones, etc.) es posible usando los sistemas integrados de la invención. Más recientemente, se han desarrollado enfoques microfluídicos para la manipulación de reactivos, por ejemplo, por Caliper Technologies (Mountain View, CA) que pueden proporcionar métodos de ensayo de microfluidos de muy alto rendimiento.

[0272] 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 suelen automatizar procedimientos completos, incluidas todas las pipetas de muestras y reactivos, dispensación de líquidos, incubaciones cronometradas y lecturas finales de la microplaca en los detectores apropiados para el ensayo. Estos sistemas configurables proporcionan un alto rendimiento y un arranque rápido, así como un alto grado de flexibilidad y personalización.

[0273] Los fabricantes de tales sistemas proporcionan protocolos detallados para varios ensayos de alto rendimiento de cribado. Así, por ejemplo, Zymark Corp. proporciona boletines técnicos que describen sistemas de detección para detectar la modulación de la transcripción génica, la unión de ligandos y similares.

[0274] Se encuentra disponible una variedad de equipos y software periféricos disponibles comercialmente para digitalizar, almacenar y analizar un video digitalizado o imágenes ópticas digitalizadas u otras imágenes de ensayo, por ejemplo, usando una PC (Intel x86 o MAC OS compatible con chip pentium, familia WINDOWS™, o computadoras basadas en UNIX (p. ej., estación de trabajo SUN™).

[0275] Los sistemas de análisis típicamente incluyen una computadora digital específicamente programada para realizar algoritmos especializados utilizando software para dirigir uno o más pasos de uno o más de los métodos aquí mencionados, y, opcionalmente, también incluyen, por ejemplo, un software de control de plataforma de secuenciación de próxima generación, software de control de líquidos de alto rendimiento, software de análisis de imágenes, software de interpretación de datos, una armadura robótica de control de líquidos para transferir soluciones desde una fuente a un destino operativamente vinculado a la computadora digital, un dispositivo de entrada (por ejemplo, un teclado de computadora) para ingresar datos a la computadora digital para controlar las operaciones o la transferencia de líquidos de alto rendimiento por el robot armadura de control de líquido tic y, opcionalmente, un escáner de imágenes para digitalizar señales de etiquetas de componentes de ensayo etiquetados. El escáner de imágenes puede interactuar con el software de análisis de imágenes para proporcionar una medición de la intensidad de la etiqueta de la sonda. Típicamente, la medición de la intensidad de la etiqueta de la sonda es interpretada por el software de interpretación de datos para mostrar si la sonda marcada se hibrida con el ADN en el soporte sólido.

[0276] En algunas realizaciones, las células, placas virales, esporas o similares, que comprenden productos de recombinación mediados por oligonucleótidos *in vitro* o realizaciones físicas de ácidos nucleicos recombinados *in silico,* pueden separarse en medios sólidos para producir colonias (o placas) individuales. Usando un selector automático de colonias (p. ej., Q-bot, Genetix, Reino Unido), se identifican colonias o placas, y se inoculan hasta 10.000 mutantes diferentes en placas de microtitulación de 96 pocillos que contienen dos bolas de vidrio de 3 mm/pocillo. El Q-bot no elige una colonia completa, sino que inserta un alfiler a través del centro de la colonia y sale con una pequeña muestra de células (o micelios) y esporas (o virus en aplicaciones de placa). El tiempo que el pin está en la colonia, el número de inmersiones para inocular el medio de cultivo, y el tiempo que el pin está en ese medio, cada tamaño de inóculo de efecto, y cada parámetro puede ser controlado y optimizado.

[0277] El proceso uniforme de selección automática de colonias, como el Q-bot, disminuye el error de manejo humano y aumenta la tasa de establecimiento de cultivos (aproximadamente 10.000/4 horas). Estos cultivos se agitan opcionalmente en una incubadora con temperatura y humedad controladas. Las bolas de vidrio opcionales en las placas de microtitulación actúan para promover la aireación uniforme de las células y la dispersión de fragmentos celulares (por ejemplo, miceliales) similares a las cuchillas de un fermentador. Los clones de cultivos de interés pueden aislarse mediante dilución limitante. Como también se describió anteriormente, las placas o las células que constituyen bibliotecas también pueden seleccionarse directamente para la producción de proteínas, ya sea mediante la detección de hibridación, actividad de proteínas, unión de proteínas a anticuerpos o similares. Para aumentar las posibilidades de identificar un grupo de tamaño suficiente, se puede utilizar una pantalla previa que aumenta el número de mutantes procesados en 10 veces. El objetivo de la detección primaria es identificar rápidamente mutantes que tengan títulos de producto iguales o mejores que la(s) cepa(s) original(es) y mover solo estos mutantes hacia el cultivo celular líquido para su posterior análisis.

[0278] Un enfoque para seleccionar diversas bibliotecas es usar un procedimiento de fase sólida masivamente paralelo para seleccionar células que expresan variantes de polinucleótidos, por ejemplo, polinucleótidos que codifican variantes de enzimas. Se encuentran disponibles aparatos de cribado de fase sólida masivamente paralelos que

utilizan absorción, fluorescencia o FRET. Ver, por ejemplo, la patente de EE.UU. 5.914.245 de Bylina, et al. (1999); ver también, http://www|.|kairos-scientific.com/; Youvan y col. (1999) "Fluorescence Imaging MicroSpectrophotometer (FIMS)" Biotechnology et alia, <www|.|et-al.com> 1:1-16; Yang et al. (1998) "High Resolution Imaging Microscope (HIRIM)" Biotechnology et alia, <www|.|et-al.com> 4:1-20; and Youvan et al. (1999) "Calibration of Fluorescence Resonance Energy Transfer in Microscopy Using Genetically Engineered GFP Derivatives on Nickel Chelating Beads" posted at www|.|kairos-scientific.com. Después del cribado mediante estas técnicas, las moléculas de interés se aíslan típicamente, y opcionalmente se secuencian usando métodos que son conocidos en la técnica. La información de la secuencia se usa como se establece en el presente documento para diseñar una nueva biblioteca de variantes de proteínas.

[0279] De manera similar, también se han desarrollado una serie de sistemas robóticos bien conocidos para las químicas de fase de solución útiles en sistemas de ensayo. Estos sistemas incluyen estaciones de trabajo automatizadas como el aparato de síntesis automatizado 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 manuales sintéticas realizadas por un científico. Cualquiera de los dispositivos anteriores es adecuado para su uso con la presente invención, por ejemplo, para el cribado de alto rendimiento de moléculas codificadas por ácidos nucleicos evolucionados como se describe aquí. La naturaleza y la implementación de modificaciones a estos dispositivos (si existen) para que puedan operar como se discute en el presente documento serán evidentes para los expertos en la materia.

X. APARATOS Y SISTEMAS DIGITALES

[0280] Como debería ser evidente, las realizaciones descritas en este documento emplean procesos que actúan bajo el control de instrucciones y/o datos almacenados o transferidos a través de uno o más sistemas informáticos. Las realizaciones descritas en el presente documento también se refieren a aparatos para realizar estas operaciones. En algunas realizaciones, el aparato está especialmente diseñado y/o construido para los propósitos requeridos, o puede ser una computadora de propósito general activada o reconfigurada selectivamente por un programa de computadora y/o estructura de datos almacenados en la computadora. Los procesos proporcionados por la presente divulgación no están inherentemente relacionados con ninguna computadora en particular u otro aparato específico. En particular, varias máquinas de uso general encuentran uso con programas escritos de acuerdo con las enseñanzas de este documento. Sin embargo, en algunas realizaciones, se construye un aparato especializado para realizar las operaciones de método requeridas. A continuación se describe una realización de una estructura particular para una variedad de estas máquinas.

[0281] Además, ciertas realizaciones de la presente divulgación se refieren a medios legibles por computadora o productos de programas de computadora que incluyen instrucciones de programa y/o datos (incluyendo estructuras de datos) para realizar diversas operaciones implementadas por computadora. Los ejemplos de medios legibles por computadora incluyen, entre otros, medios magnéticos como discos duros; medios ópticos tales como dispositivos de CD-ROM y dispositivos holográficos; medios magnetoópticos; y dispositivos de memoria de semiconductores como la memoria flash. Los dispositivos de hardware como los dispositivos de memoria de solo lectura (ROM) y los dispositivos de memoria de acceso aleatorio (RAM) pueden configurarse para almacenar instrucciones del programa. Los dispositivos de hardware como los circuitos integrados específicos de la aplicación (ASIC) y los dispositivos lógicos programables (PLD) se pueden configurar para almacenar las instrucciones del programa y ejecutarlas. No se pretende que la presente divulgación se limite a ningún medio legible por computadora en particular o cualquier otro producto de programa de computadora que incluya instrucciones y/o datos para realizar operaciones implementadas por computadora.

[0282] Los ejemplos de instrucciones del programa incluyen, entre otros, códigos de bajo nivel, como los producidos por un compilador, y archivos que contienen código de nivel superior que la computadora puede ejecutar utilizando un intérprete. Además, las instrucciones del programa incluyen, entre otras, el código de máquina, el código fuente y cualquier otro código que controle directa o indirectamente el funcionamiento de una máquina informática de acuerdo con la presente divulgación. El código puede especificar entradas, salidas, cálculos, condicionales, ramificaciones, bucles iterativos, etc.

[0283] En un ejemplo ilustrativo, los métodos de incorporación de códigos descritos en este documento están incorporados en un medio fijo o componente de programa transmisible que contiene instrucciones lógicas y/o datos que cuando se carga en un dispositivo informático configurado adecuadamente, hace que el dispositivo realice una operación genética simulada (GO) en una o más cadenas de caracteres. La Figura 4 muestra un ejemplo de dispositivo digital 800 que es un aparato lógico que puede leer instrucciones del medio 817, el puerto de red 819, el teclado de entrada del usuario 809, la entrada del usuario 811 u otros medios de entrada. Posteriormente, el aparato 800 puede usar esas instrucciones para dirigir operaciones estadísticas en el espacio de datos, por ejemplo, para construir uno o más conjuntos de datos (por ejemplo, para determinar una pluralidad de miembros representativos del espacio de datos). Un tipo de aparato lógico que puede incorporar realizaciones reveladas es un sistema informático como en el sistema informático 800 que comprende la CPU 807, el teclado opcional de dispositivos de entrada de usuario 809 y el dispositivo señalador GUI 811, así como componentes periféricos tales como unidades de disco 815 y monitor 805 (que muestra cadenas de caracteres modificadas GO y proporciona una selección simplificada de subconjuntos de

tales cadenas de caracteres por un usuario. El medio fijo 817 se usa opcionalmente para programar el sistema general y puede incluir, por ejemplo, un medio óptico o magnético de tipo disco u otra memoria electrónica elemento de almacenamiento. El puerto de comunicación 819 puede usarse para programar el sistema y puede representar cualquier tipo de conexión de comunicación.

[0284] Ciertas realizaciones también pueden realizarse dentro de los circuitos de un circuito integrado específico de aplicación (ASIC) o dispositivo lógico programable (PLD). En tal caso, las realizaciones se implementan en un lenguaje descriptor legible por computadora que puede usarse para crear un ASIC o PLD. Algunas realizaciones de la presente divulgación se implementan dentro de los circuitos o procesadores lógicos de una variedad de otros aparatos digitales, tales como PDA, sistemas de computadoras portátiles, pantallas, equipos de edición de imágenes, etc.

[0285] En algunas realizaciones, la presente divulgación se refiere a un producto de programa de computadora que comprende uno o más medios de almacenamiento legibles por computadora que han almacenado en él instrucciones ejecutables por computadora que, cuando son ejecutadas por uno o más procesadores de un sistema informático, hacen que el sistema informático implemente un método para el cribado virtual de variantes de proteínas y/o evolución dirigida *in silico* de proteínas que tienen la actividad deseada. Tal método puede ser cualquier método descrito en este documento, como los abarcados por las Figuras y el pseudocódigo. En algunas realizaciones, por ejemplo, el método recibe datos de secuencia para una pluralidad de enzimas, crea modelos de homología tridimensional de moléculas biológicas, acopla los modelos de homología de enzimas con una o más representaciones computacionales de sustratos y deriva datos estructurales con respecto a parámetros geométricos. con referencia a las enzimas y sustratos. En algunas realizaciones, el método puede desarrollar adicionalmente modelos de actividad de secuencia filtrando datos con referencia a los datos estructurales modelados. Las bibliotecas de variantes se pueden usar en la evolución dirigida reiterativa, lo que puede dar como resultado enzimas de propiedades beneficiosas deseadas.

[0286] En algunas realizaciones, el acoplamiento de los modelos de homología de enzimas con una o más representaciones computacionales de sustratos se realiza mediante un programa de acoplamiento en un sistema informático que utiliza una representación computacional de un ligando y representaciones computacionales de los sitios de unión de una pluralidad de variantes de las maneras descritas en este documento. En diversas realizaciones, el programa de acoplamiento evalúa la energía de unión entre una postura del sustrato y la enzima. Para una variante de proteína que se acopla con éxito con el ligando, el sistema determina valores geométricos con respecto al ligando y la proteína participantes. En diversas realizaciones, el sistema informático construye un modelo de actividad de secuencia entrenando una máquina de vectores de soporte. En diversas realizaciones, el sistema informático usa algoritmos genéticos para filtrar datos no informativos, proporcionando así un subconjunto de datos para entrenar la máquina de vectores de soporte.

XI. REALIZACIONES EN SITIOS WEB Y COMPUTACIÓN EN LA NUBE

[0287] Internet incluye computadoras, dispositivos de información y redes de computadoras que están interconectadas a través de enlaces de comunicación. Las computadoras interconectadas intercambian información utilizando diversos servicios, como correo electrónico, ftp, la World Wide Web ("WWW") y otros servicios, incluidos servicios seguros. Se puede entender que el servicio WWW permite que un sistema informático de servidor (por ejemplo, un servidor web o un sitio web) envíe páginas web de información a un dispositivo de información de cliente remoto o sistema informático. El sistema informático del cliente remoto puede mostrar las páginas web. En general, cada recurso (p. ej., computadora o página web) de la WWW es identificable de manera única por un Localizador Uniforme de Recursos ("URL"). Para ver o interactuar con una página web específica, un sistema informático cliente especifica una URL para esa página web en una solicitud. La solicitud se reenvía a un servidor que admite esa página web. Cuando el servidor recibe la solicitud, envía esa página web al sistema de información del cliente. Cuando el sistema informático del cliente recibe esa página web, puede mostrar la página web utilizando un navegador o puede interactuar con la página web o la interfaz según lo dispuesto. Un navegador es un módulo lógico que efectúa la solicitud de páginas web y muestra o interactúa con las páginas web.

[0288] Actualmente, las páginas web visualizables se definen típicamente usando un lenguaje de marcado de hipertexto ("HTML"). HTML proporciona un conjunto estándar de etiquetas que definen cómo se mostrará una página web. Un documento HTML contiene varias etiquetas que controlan la visualización de texto, gráficos, controles y otras características. El documento HTML puede contener URL de otras páginas web disponibles en ese sistema informático del servidor u otros sistemas informáticos del servidor. Las URL también pueden indicar otros tipos de interfaces, incluidas las secuencias de comandos CGI o las interfaces ejecutables, que los dispositivos de información utilizan para comunicarse con servidores o dispositivos de información remotos sin necesariamente mostrar información a un usuario.

[0289] Internet es especialmente propicio para proporcionar servicios de información a uno o más clientes remotos. Los servicios pueden incluir artículos (por ejemplo, cotizaciones de música o acciones) que se entregan electrónicamente a un comprador a través de Internet. Los servicios también pueden incluir el manejo de pedidos de artículos (por ejemplo, comestibles, libros o compuestos químicos o biológicos, etc.) que pueden entregarse a través de canales de distribución convencionales (por ejemplo, un transportista común). Los servicios también pueden incluir el manejo de pedidos de artículos, como reservas de aerolíneas o teatros, a los que el comprador accede en un

momento posterior. Un sistema de computadora servidor puede proporcionar una versión electrónica de una interfaz que enumera los elementos o servicios que están disponibles. Un usuario o un comprador potencial puede acceder a la interfaz utilizando un navegador y seleccionar varios elementos de interés. Cuando el usuario ha completado la selección de los elementos deseados, el sistema informático del servidor puede solicitar al usuario la información necesaria para completar el servicio. Esta información de pedido específica de la transacción puede incluir el nombre del comprador u otra identificación, una identificación para el pago (como un número de orden de compra corporativa o un número de cuenta) o información adicional necesaria para completar el servicio, como información de vuelo.

[0290] Entre los servicios de particular interés que se pueden proporcionar a través de Internet y a través de otras redes se encuentran los datos biológicos y las bases de datos biológicas. Dichos servicios incluyen una variedad de servicios proporcionados por el Centro Nacional de Información Biotecnológica (NCBI) de los Institutos Nacionales de Salud (NIH). NCBI se encarga de crear sistemas automatizados para almacenar y analizar conocimientos sobre biología molecular, bioquímica y genética; facilitando el uso de tales bases de datos y software por la comunidad médica y de investigación; coordinar esfuerzos para recopilar información biotecnológica tanto a nivel nacional como internacional; y realizar investigaciones sobre métodos avanzados de procesamiento de información por computadora para analizar la estructura y función de moléculas biológicamente importantes.

[0291] NCBI es responsable de la base de datos de secuencias de ADN GenBank®. La base de datos ha sido construida a partir de secuencias enviadas por laboratorios individuales y por 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 secuencia de patentes enviados a la Oficina de Patentes y Marcas de EE.UU. Además de GenBank®, NCBI apoya y distribuye una variedad de bases de datos para las comunidades médicas y científicas. Estos incluyen la herencia mendeliana en línea en el hombre (OMIM), la base de datos de modelado molecular (MMDB) de estructuras de proteínas 3D, la colección única de secuencias de genes humanos (UniGene), un mapa genético del genoma humano, el buscador de taxonomía y el proyecto de anatomía de genoma del cáncer (CGAP), en colaboración con el Instituto Nacional del Cáncer. Entrez es el sistema de búsqueda y recuperación de NCBI que brinda a los usuarios acceso integrado a secuencia, mapeo, taxonomía y datos estructurales. Entrez también proporciona vistas gráficas de secuencias y mapas cromosómicos. Una característica de Entrez es la capacidad de recuperar secuencias, estructuras y referencias relacionadas. BLAST, como se describe aquí, es un programa de búsqueda de similitud de secuencia desarrollado en NCBI para identificar genes y características genéticas que pueden ejecutar búsquedas de secuencia en toda la base de datos de ADN. Entre las herramientas de software adicionales proporcionadas por NCBI se incluyen: Open Reading Frame Finder (ORF Finder), PCR electrónica y las herramientas de envío de secuencias, Sequin y BankIt. Las diversas bases de datos y herramientas de software de NCBI están disponibles en WWW, FTP o servidores de correo electrónico. Hay más información disponible en www|.|ncbi.nlm.nih.gov.

[0292] Algunos datos biológicos disponibles en Internet son datos que generalmente se ven con un "complemento" de navegador especial u otro código ejecutable. Un ejemplo de dicho sistema es CHIME, un complemento de navegador que permite una visualización tridimensional virtual interactiva de estructuras moleculares, incluidas las estructuras moleculares biológicas. Más información sobre CHIME está disponible en www|.|mdlchime.com/chime/.

[0293] Una variedad de empresas e instituciones proporcionan sistemas en línea para ordenar compuestos biológicos. Se pueden encontrar ejemplos de tales sistemas en www|.|genosys.com/ oligo_custinfo.cfin o www|.|genomictechnologies.com/Qbrowser2_FP.htm1. Típicamente, estos sistemas aceptan algún descriptor de un Compuesto Biológico deseado (como un oligonucleótido, cadena de ADN, cadena de ARN, secuencia de aminoácidos, etc.) y luego el compuesto solicitado se fabrica y se envía al cliente en una solución líquida u otra forma apropiada.

[0294] Como los métodos proporcionados en este documento pueden implementarse en un sitio web como se describe más adelante, los resultados computacionales o físicos que involucran polipéptidos o polinucleótidos producidos por algunas realizaciones de la divulgación pueden proporcionarse a través de Internet de manera similar a la información biológica y compuestos descritos anteriormente.

[0295] Para ilustrar adicionalmente, los métodos de esta invención pueden implementarse en un entorno informático localizado o distribuido. En un entorno distribuido, los métodos pueden implementarse en una sola computadora que comprende múltiples procesadores o en una multiplicidad de computadoras. Las computadoras se pueden vincular, por ejemplo, a través de un bus común, pero más preferiblemente las computadoras son nodos en una red. La red puede ser una red local o de área extensa generalizada o dedicada y, en ciertas realizaciones preferidas, las computadoras pueden ser componentes de una Intranet o de Internet.

[0296] En una realización de Internet, un sistema cliente normalmente ejecuta un navegador web y está acoplado a una computadora servidor que ejecuta un servidor web. El navegador web suele ser un programa como el Web Explorer de IBM, el explorador de Internet de Microsoft, NetScape, Opera o Mosaic. El servidor web es típicamente, pero no necesariamente, un programa como el HTTP Daemon de IBM u otro www daemon (por ejemplo, formas del programa basadas en LINUX). La computadora del cliente está acoplada bidireccionalmente con la computadora del servidor a través de una línea o mediante un sistema inalámbrico. A su vez, la computadora servidor está acoplada bidireccionalmente con un sitio web (servidor que aloja el sitio web) que proporciona acceso al software que

implementa los métodos de esta invención.

[0297] Como se mencionó, un usuario de un cliente conectado a la Intranet o Internet puede hacer que el cliente solicite recursos que son parte de los sitios web que alojan las aplicaciones que proporcionan una implementación de los métodos de esta invención. Los programas del servidor luego procesan la solicitud para devolver los recursos especificados (suponiendo que estén actualmente disponibles). La convención de nomenclatura estándar (es decir, localizador uniforme de recursos ("URL")) abarca varios tipos de nombres de ubicaciones, que actualmente incluyen subclases como el protocolo de transporte de hipertexto ("http"), el protocolo de transporte de archivos ("ftp"), gopher y servicio de información de área amplia ("WAIS"). Cuando se descarga un recurso, puede incluir las URL de recursos adicionales. Por lo tanto, el usuario del cliente puede aprender fácilmente sobre la existencia de nuevos recursos que no había solicitado específicamente.

[0298] El software que implementa el (los) método(s) de esta invención se puede ejecutar de forma local en el servidor que aloja el sitio web en una arquitectura cliente-servidor cierta. Por lo tanto, la computadora cliente envía solicitudes al servidor host que ejecuta los procesos solicitados localmente y luego descarga los resultados nuevamente al cliente. Alternativamente, los métodos de esta invención pueden implementarse en un formato de "múltiples niveles" en donde el cliente realiza localmente un componente de los métodos. Esto puede implementarse mediante software descargado del servidor a solicitud del cliente (por ejemplo, una aplicación Java) o puede implementarse mediante software instalado "permanentemente" en el cliente.

[0299] En una realización, las aplicaciones que implementan los métodos de esta invención se dividen en tramas. En este paradigma, es útil ver una aplicación no tanto como una colección de características o funcionalidades, sino como una colección de marcos o vistas discretas. Una aplicación típica, por ejemplo, generalmente incluye un conjunto de elementos de menú, cada uno de los cuales invoca un marco particular, es decir, un formulario que manifiesta cierta funcionalidad de la aplicación. Con esta perspectiva, una aplicación se ve no como un cuerpo monolítico de código sino como una colección de applets o paquetes de funcionalidades. De esta manera, desde un navegador, un usuario seleccionaría un enlace de página web que, a su vez, invocaría un marco particular de la aplicación (es decir, una sub-aplicación). Así, por ejemplo, uno o más cuadros pueden proporcionar funcionalidad para ingresar y/o codificar molécula(s) biológica(s) en uno o más espacios de datos, mientras que otro cuadro proporciona herramientas para refinar un modelo del espacio de datos.

[**0300]** En ciertas realizaciones, los métodos de esta invención se implementan como uno o más cuadros que proporcionan, por ejemplo, las siguientes funcionalidades: funciones para codificar dos o más moléculas biológicas en cadenas de caracteres para proporcionar una colección de dos o más cadenas de caracteres iniciales diferentes en las que cada una de dichas moléculas biológicas comprende un conjunto seleccionado de subunidades; funciones para seleccionar al menos dos subcadenas de las cadenas de caracteres; funciones para concatenar las subcadenas para formar una o más cadenas de productos de aproximadamente la misma longitud que una o más de las cadenas de caracteres iniciales; funciones para agregar (colocar) las cadenas de productos a una colección de cadenas; funciones para crear y manipular representación/modelos computacionales de enzimas y sustratos, funciones para acoplar una representación computacional de un sustrato (por ejemplo, un ligando) con la representación computacional de una enzima (por ejemplo, una proteína); funciones para aplicar dinámica molecular a modelos moleculares; funciones para calcular varias restricciones entre moléculas que afectan las reacciones químicas que involucran a las moléculas (por ejemplo, distancia o ángulo entre un resto de sustrato y un sitio activo de enzima); y funciones para implementar cualquier característica establecida en este documento.

[**0301]** Una o más de estas funcionalidades también pueden implementarse exclusivamente en un servidor o en una computadora cliente. Estas funciones, por ejemplo, funciones para crear o manipular modelos computacionales de moléculas biológicas, pueden proporcionar una o más ventanas en las que el usuario puede insertar o manipular representación(es) de moléculas biológicas. Además, las funciones también, opcionalmente, proporcionan acceso a bases de datos privadas y/o públicas accesibles a través de una red local y/o la intranet, por lo que una o más secuencias contenidas en las bases de datos pueden introducirse en los métodos de esta invención. Así, por ejemplo, en una realización, el usuario puede, opcionalmente, tener la capacidad de solicitar una búsqueda de GenBank® e ingresar una o más de las secuencias devueltas por dicha búsqueda en una función de codificación y/o generación de diversidad.

[**0302]** Los métodos de implementación de Intranet y/o Intranet de procesos computacionales y/o de acceso a datos son bien conocidos por los expertos en la materia y están documentados en gran detalle (ver, por ejemplo, Cluer et al. (1992) "A General Framework for the Optimization of Object-Oriented Queries," Proc SIGMOD International Conference on Management of Data, San Diego, California, Jun. 2-5, 1992,, Registro SIGMOD, vol. 21, número 2, junio de 1992; Stonebraker, M., Editor; ACM Press, págs. 383-392; ISO-ANSI, Borrador de trabajo, " Information Technology-Database Language SQL," Jim Melton, Editor, International Organization for Standardization and American National Standards Institute, Jul. 1992; Microsoft Corporation, "ODBC 2.0 Programmer's Reference and SDK Guide. The Microsoft Open Database Standard for Microsoft Windows.™ and Windows NT™, Microsoft Open Database Connectivity.TM. Software Development Kit," 1992, 1993, 1994 Microsoft Press, pp. 3-30 and 41-56; ISO borrador de trabajo, "Database Language SQL-Part 2: Foundation (SQL/Foundation)," CD9075-2:199.chi.SQL, Sep. 11, 1997, y similares). En el documento WO 00/42559, titulado "METHODS OF POPULATING DATA STRUCTURES

FOR USE IN EVOLUTIONARY SIMULATIONS", Selifonov y Stemmer encuentran detalles relevantes adicionales con respecto a las aplicaciones basadas en la web.

[**0303]** En algunas realizaciones, los métodos para explorar, seleccionar y/o desarrollar secuencias de polinucleótidos o polipéptidos pueden implementarse como un sistema multiusuario en un sistema informático con una pluralidad de unidades de procesamiento y memorias distribuidas a través de una red informática, en donde la red puede incluir intranet en LAN y/o Internet. En algunas realizaciones, la arquitectura informática distribuida implica una "nube", que es una colección de sistemas informáticos disponibles a través de una red informática para el cálculo y el almacenamiento de datos. El entorno informático que implica una nube se denomina entorno informático en la nube. En algunas realizaciones, uno o más usuarios pueden acceder a las computadoras de la nube distribuidas a través de una intranet y/o Internet. En algunas realizaciones, un usuario puede acceder de forma remota, a través de un cliente web, a computadoras de servidor que implementan los métodos para seleccionar y/o desarrollar variantes de proteínas descritas anteriormente.

[**0304]** En algunas realizaciones que implican un entorno de computación en la nube, las máquinas virtuales (VM) se aprovisionan en las computadoras del servidor, y los resultados de las máquinas virtuales se pueden enviar de vuelta al usuario. Una máquina virtual (VM) es una emulación basada en software de una computadora. Las máquinas virtuales pueden estar basadas en especificaciones de una computadora hipotética o emular la arquitectura y las funciones de una computadora del mundo real. La estructura y las funciones de las máquinas virtuales son bien conocidas en la técnica. Por lo general, una VM se instala en una plataforma host que incluye hardware del sistema, y la VM en sí incluye hardware del sistema virtual y software invitado.

[**0305]** El hardware del sistema host para una VM incluye una o más unidades de procesamiento central (CPU), memoria, uno o más discos duros y varios otros dispositivos. El hardware del sistema virtual de la VM incluye una o más CPU virtuales, memoria virtual, uno o más discos duros virtuales y uno o más dispositivos virtuales. El software invitado de la VM incluye software de sistema invitado y aplicaciones invitadas. En algunas implementaciones, el software del sistema invitado incluye un sistema operativo invitado con controladores para dispositivos virtuales. En algunas implementaciones, las aplicaciones invitadas de la VM incluyen al menos una instancia de un sistema virtual de detección de proteínas como se describió anteriormente.

[**0306]** En algunas realizaciones, el número de máquinas virtuales aprovisionadas se puede escalar a la carga computacional del problema a resolver. En algunas realizaciones, un usuario puede solicitar una máquina virtual desde una nube, la VM incluye un sistema de detección virtual. En algunas realizaciones, el entorno de computación en la nube puede aprovisionar una VM basada en la solicitud del usuario. En algunas realizaciones, puede existir una VM en una imagen de VM previamente almacenada, que puede almacenarse en un repositorio de imágenes. El entorno de computación en la nube puede buscar y transferir la imagen a un servidor o un sistema de usuario. El entorno de computación en la nube puede iniciar la imagen en el servidor o sistema de usuario.

REIVINDICACIONES

1. Un método implementado por computadora para llevar a cabo la evolución dirigida, el método comprende:

(a) recibir un conjunto de datos sin filtrar que tiene información de mediciones físicas de moléculas, en donde el conjunto de datos sin filtrar comprende la siguiente información para cada una de una pluralidad de variantes biomoleculares: (i) actividad de la biomolécula variante en un ligando en un sitio de unión de la biomolécula variante, (ii) una secuencia de la biomolécula variante, en donde la secuencia es una secuencia de ácido nucleico o una secuencia de proteínas, y (iii) uno o más parámetros geométricos que caracterizan la geometría del ligando en el sitio de unión de la biomolécula variante;
(b) filtrar el conjunto de datos sin filtrar para producir un subconjunto de datos filtrados, eliminando información para una o más de las variantes biomoleculares, en donde el filtrado elimina al menos uno de los parámetros geométricos del conjunto de datos sin filtrar y/o elimina del conjunto de datos sin filtrar ciertas variantes biomoleculares que tienen valores de parámetros geométricos fuera de los rangos definidos, y en donde el filtrado comprende modelos de actividad de secuencia de entrenamiento con una pluralidad de subconjuntos de datos seleccionados y prueba de la capacidad de los modelos de actividad de secuencia entrenados con la pluralidad de subconjuntos de datos seleccionados para predecir la actividad de una biomolécula variante en el ligando en el sitio de unión de la biomolécula variante en función de variables independientes y, por lo tanto, identifica un subconjunto de datos filtrados que proporciona un modelo de actividad de secuencia con mayor capacidad para predecir la actividad de una biomolécula variante en el ligando en el sitio de unión de la biomolécula variante en función de variables independientes que un modelo de actividad de secuencia entrenado con el conjunto de datos sin filtrar, en donde la secuencia de la biomolécula variante y los parámetros geométricos filtrados que caracterizan la geometría del ligando en el sitio de unión de la biomolécula variante son variables independientes en los modelos de actividad de secuencia entrenados con la pluralidad de subconjuntos de datos seleccionados, cada subconjunto de datos seleccionado tiene información para al menos uno de los parámetros geométricos y/o ciertas variantes biomoleculares que tienen valores de parámetros geométricos fuera de los rangos definidos eliminados del conjunto de datos sin filtrar de (a); y
(c) aplicar un modelo de actividad de secuencia entrenado usando el subconjunto de datos filtrados para identificar una o más variantes de biomolécula nuevas predichas para tener actividad que cumpla con uno o más criterios, en donde cada una de las una o más variantes de biomolécula nueva tiene una secuencia de ácido nucleico o proteína que difiere de las secuencias de las variantes de biomoléculas que proporcionan información para el conjunto de datos sin filtrar de (a).

2. El método de la reivindicación 1, en donde el filtrado del conjunto de datos se realiza con un algoritmo genético, opcionalmente en donde el algoritmo genético varía los umbrales para eliminar la información asociada con los parámetros geométricos para una o más de las variantes biomoleculares.

3. El método de la reivindicación 1 o 2, en donde aplicar el modelo de actividad de secuencia entrenado usando el subconjunto de datos filtrados para identificar una o más variantes de biomolécula nueva comprende realizar un algoritmo genético en donde se evalúan nuevas variantes de biomolécula potenciales usando el modelo de actividad de secuencia entrenado usando el subconjunto de datos filtrados como una función de aptitud.

4. El método de la reivindicación 1 o 2, que comprende

además producir un modelo estructural para cada una de las nuevas variantes de biomolécula; y
el uso de modelos estructurales para generar parámetros geométricos que caracterizan la geometría del ligando en los sitios de unión de las nuevas variantes de biomoléculas.

5. El método de cualquiera de las reivindicaciones anteriores, que comprende además recibir modelos estructurales de variantes de biomoléculas y determinar el uno o más parámetros geométricos usando los modelos estructurales.

6. El método de la reivindicación 65, en donde los modelos estructurales son modelos de homología, y opcionalmente en donde los modelos de homología se preparan usando detalles de medición estructural física de biomoléculas.

7. El método de la reivindicación 6, en donde los detalles de medición estructural física de las biomoléculas comprenden posiciones tridimensionales de átomos obtenidos por RMN o cristalografía de rayos X.

8. El método de la reivindicación 5, que comprende además usar un acoplador para determinar el uno o más parámetros geométricos.

9. El método de cualquiera de las reivindicaciones anteriores, en donde la información para cada una de una pluralidad de variantes biomoleculares comprende además (iv) una energía de interacción que caracteriza la interacción del ligando en el sitio de unión.

10. El método de la reivindicación 9, que comprende además usar un acoplador para determinar la energía de

interacción.

11. El método de cualquiera de las reivindicaciones anteriores, en donde el modelo de actividad de secuencia entrenado usando el subconjunto de datos filtrados se obtiene mediante una máquina de vectores de soporte, una regresión lineal múltiple, una regresión de componentes principales, una regresión parcial de mínimos cuadrados o una red neuronal.

12. El método de cualquiera de las reivindicaciones anteriores, en donde la pluralidad de variantes biomoleculares comprende una pluralidad de enzimas, y opcionalmente en donde la actividad de la biomolécula variante en un ligando es la actividad de una enzima en un sustrato.

13. El método de una cualquiera de las reivindicaciones anteriores, que comprende además sintetizar las nuevas variantes de biomolécula y opcionalmente analizar la actividad de las nuevas variantes de biomolécula.

14. Un producto de programa de computadora que comprende uno o más medios de almacenamiento no transitorios legibles por computadora que han almacenado en él instrucciones ejecutables por computadora que, cuando son ejecutadas por uno o más procesadores de un sistema informático, hacen que el sistema informático implemente un método de acuerdo con cualquiera de las reivindicaciones 1-12.

15. Un sistema informático que comprende:

uno o más procesadores; memoria del sistema; y uno o más medios de almacenamiento legibles por computadora que han almacenado en él instrucciones ejecutables por computadora que, cuando son ejecutadas por uno o más procesadores, hacen que el sistema informático implemente un método para llevar a cabo la evolución dirigida de acuerdo con cualquiera de las reivindicaciones 1-12.

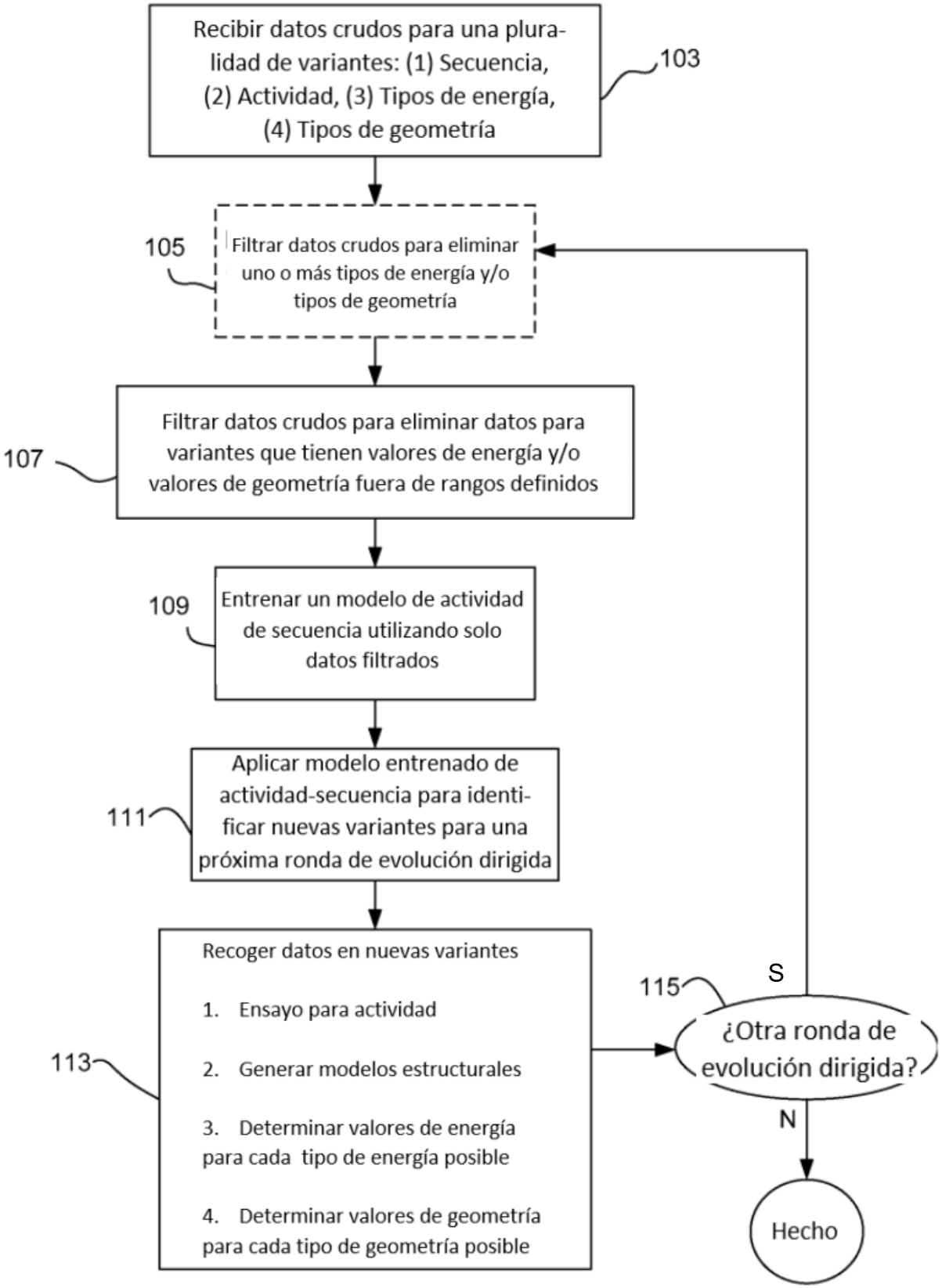
Recibir datos crudos para una pluralidad de variantes: (1) Secuencia, (2) Actividad, (3) Tipos de energía, (4) Tipos de geometría
103
105
Filtrar datos crudos para eliminar uno o más tipos de energía y/o tipos de geometría
Filtrar datos crudos para eliminar datos para variantes que tienen valores de energía y/o valores de geometría fuera de rangos definidos
107
109
Entrenar un modelo de actividad de secuencia utilizando solo datos filtrados
Aplicar modelo entrenado de actividad-secuencia para identificar nuevas variantes para una próxima ronda de evolución dirigida
111
Recoger datos en nuevas variantes
1. Ensayo para actividad
2. Generar modelos estructurales
3. Determinar valores de energía para cada tipo de energía posible
4. Determinar valores de geometría para cada tipo de geometría posible
113
115
S
¿Otra ronda de evolución dirigida?
N
Hecho

FIG. 1A

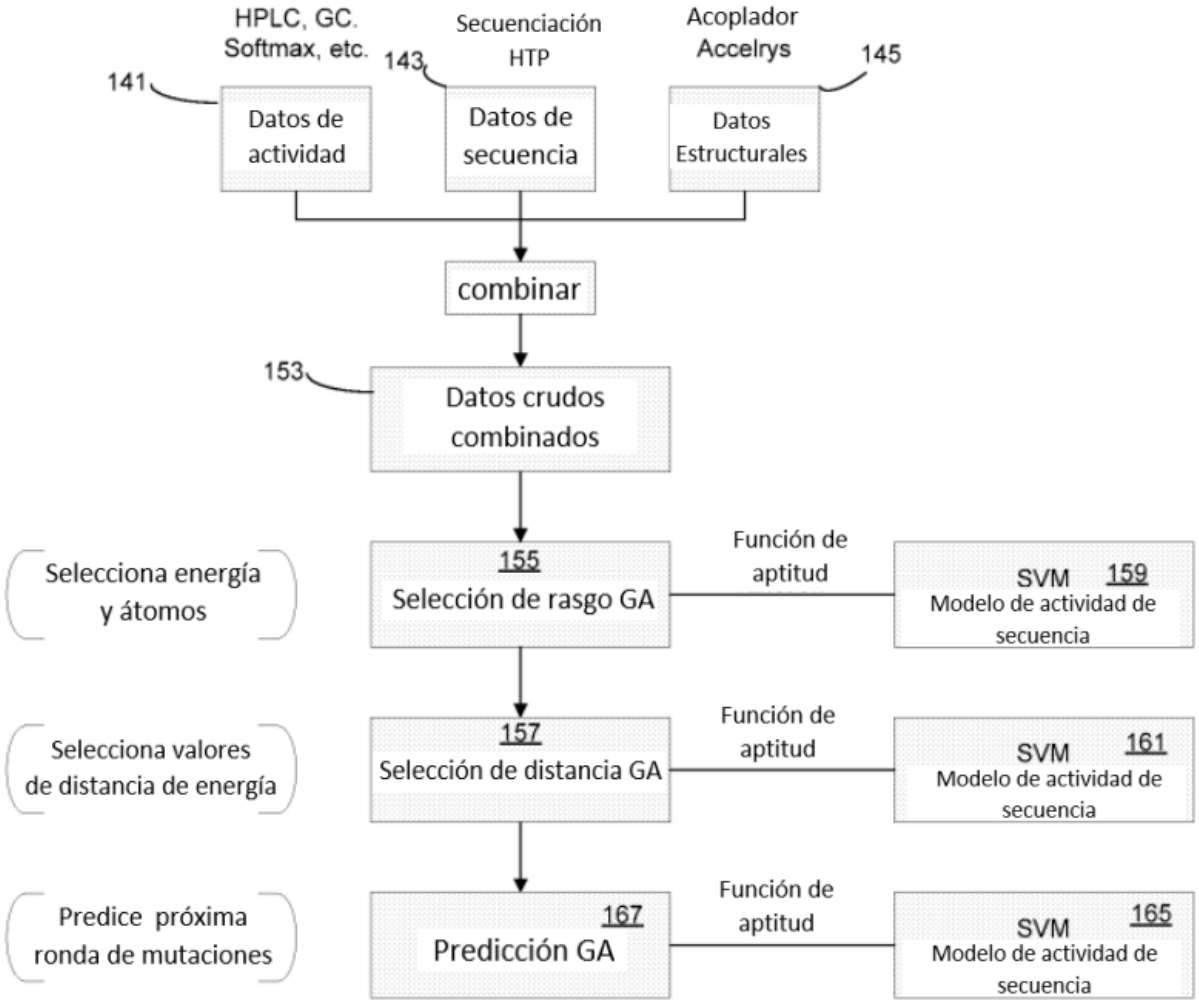
HPLC, GC.
Softmax, etc.
Secuenciación
HTP
Acoplador
Accelrys
141
143
145
Datos de actividad
Datos de secuencia
Datos Estructurales
combinar
153
Datos crudos combinados
Selecciona energía y átomos
155
Selección de rasgo GA
Función de aptitud
SVM 159
Modelo de actividad de secuencia
Selecciona valores de distancia de energía
157
Selección de distancia GA
Función de aptitud
SVM 161
Modelo de actividad de secuencia
Predice próxima ronda de mutaciones
167
Predicción GA
Función de aptitud
SVM 165
Modelo de actividad de secuencia

FIG. 1B

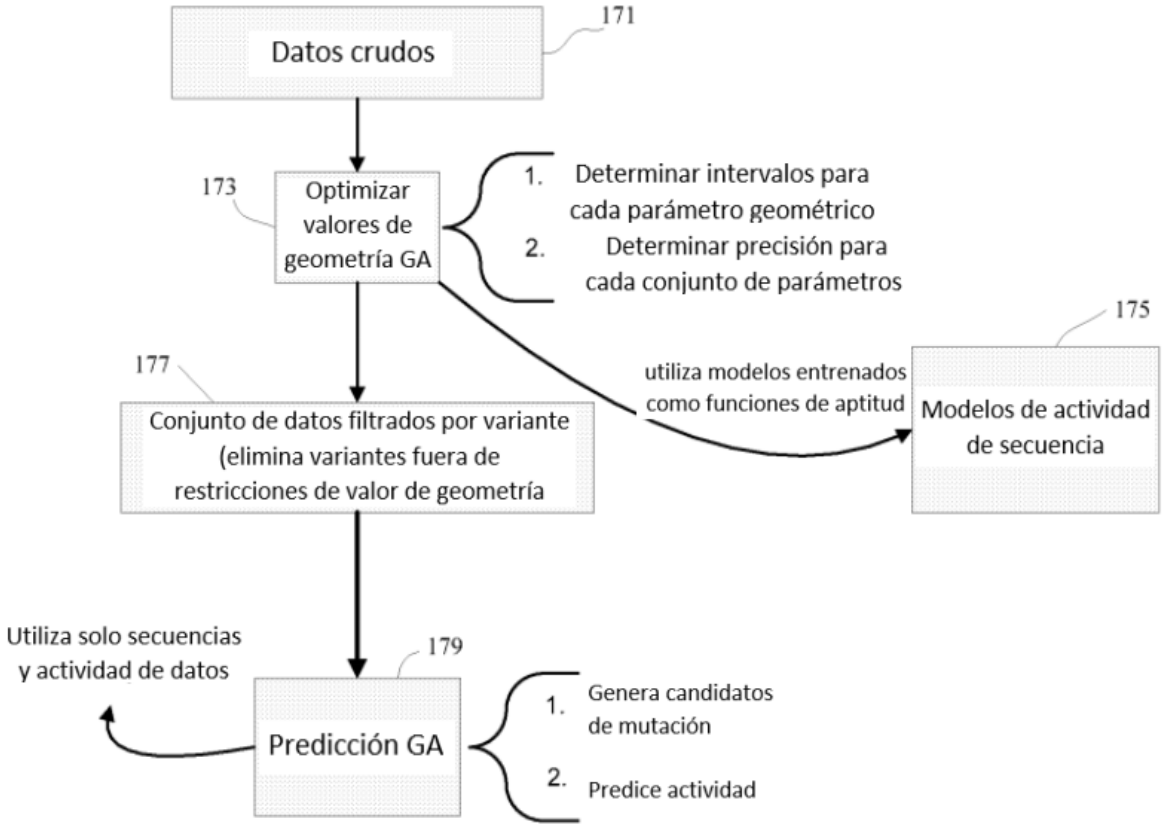
Datos crudos
171
173
Optimizar valores de geometría GA
1. Determinar intervalos para cada parámetro geométrico
2. Determinar precisión para cada conjunto de parámetros
175
utiliza modelos entrenados como funciones de aptitud
Modelos de actividad de secuencia
177
Conjunto de datos filtrados por variante (elimina variantes fuera de restricciones de valor de geometría
Utiliza solo secuencias y actividad de datos
179
Predicción GA
1. Genera candidatos de mutación
2. Predice actividad

FIG. 1C

Variante	Actividad	Secuencia			Energía (kcal/mol)		Geometría (Å)				
		P1	P2	P3	Total	Interactuar	N_1	P	$C_{(O)}$	$C_{(H3)}$	$O_{(H)}$
1	0.3	A	C	E	0.43	0.38	2.4	3.5	0.4	5.5	3.1
2	4.4	F	G	I	-2.1	-2.2	1.3	0.8	3.5	5.6	4.4
3	2.3	K	L	N	-3.1	-3.3	0.6	1.2	2.1	0.3	4.4
4	5.1	P	Q	S	-5.2	-5.6	0.3	0.7	1.1	3.1	2.1
5	1.3	T	V	Y	1.8	1.2	3.2	2.1	4.2	0.3	1.7
...	...	...	...	...	...	...	...	...	...	...	...
n	4.1	X	D	H	-4.1	-4.2	0.7	1.1	2.2	2.5	3.1

A. Datos brutos de actividad de secuencia

1 1 1 1 0 1 0

Variante	Actividad	Secuencia			Energía (kcal/mol)		Geometría (Å)				
		P1	P2	P3	Total	Interactuar	N_1	P	$C_{(O)}$	$C_{(H3)}$	$O_{(H)}$
1	0.3	A	C	E	0.43	0.38	2.4	3.5	0.4	5.5	3.1
2	4.4	F	G	I	-2.1	-2.2	1.3	0.8	3.5	5.6	4.4
3	2.3	K	L	N	-3.1	-3.3	0.6	1.2	2.1	0.3	4.4
4	5.1	P	Q	S	-5.2	-5.6	0.3	0.7	1.1	3.1	2.1
5	1.3	T	V	Y	1.8	1.2	3.2	2.1	4.2	0.3	1.7
...	...	...	...	...	...	...	...	...	...	...	...
n	4.1	X	D	H	-4.1	-4.2	0.7	1.1	2.2	2.5	3.1

B. Datos de actividad de secuencia filtrando columnas de datos

(GA individual: E Toal=1, E $_{\text{Interactuar}}$1, N_1=1, P=1, $C_{(O)}$=0, $C_{(H3)}$=1, $O_{(H)}$=0)

P>2.8

E Total>1.5

Variante	Actividad	Secuencia			Energía (kcal/mol)		Geometría (Å)				
		P1	P2	P3	Total	Interactuar	N_1	P	$C_{(O)}$	$C_{(H3)}$	$O_{(H)}$
~~1~~	~~0.3~~	~~A~~	~~C~~	~~E~~	~~0.43~~	~~0.38~~	~~2.4~~	~~3.5~~	~~0.4~~	~~5.5~~	~~3.1~~
2	4.4	F	G	I	-2.1	-2.2	1.3	0.8	3.5	5.6	4.4
3	2.3	K	L	N	-3.1	-3.3	0.6	1.2	2.1	0.3	4.4
4	5.1	P	Q	S	-5.2	-5.6	0.3	0.7	1.1	3.1	2.1
~~5~~	~~1.3~~	~~T~~	~~V~~	~~Y~~	~~1.8~~	~~1.2~~	~~3.2~~	~~2.1~~	~~4.2~~	~~0.3~~	~~1.7~~
...	...	...	...	...	...	...	...	...	...	...	...
n	4.1	X	D	H	-4.1	-4.2	0.7	1.1	2.2	2.5	3.1

C. Datos de actividad de secuencia filtrando filas de datos

(GA individual: E Toal>1.5, E $_{\text{Interactuar}}$>1.5, N_1>3.3, P>2.8, $C_{(O)}$>3.6, $C_{(H3)}$>6, $O_{(H)}$>6)

FIG. 2

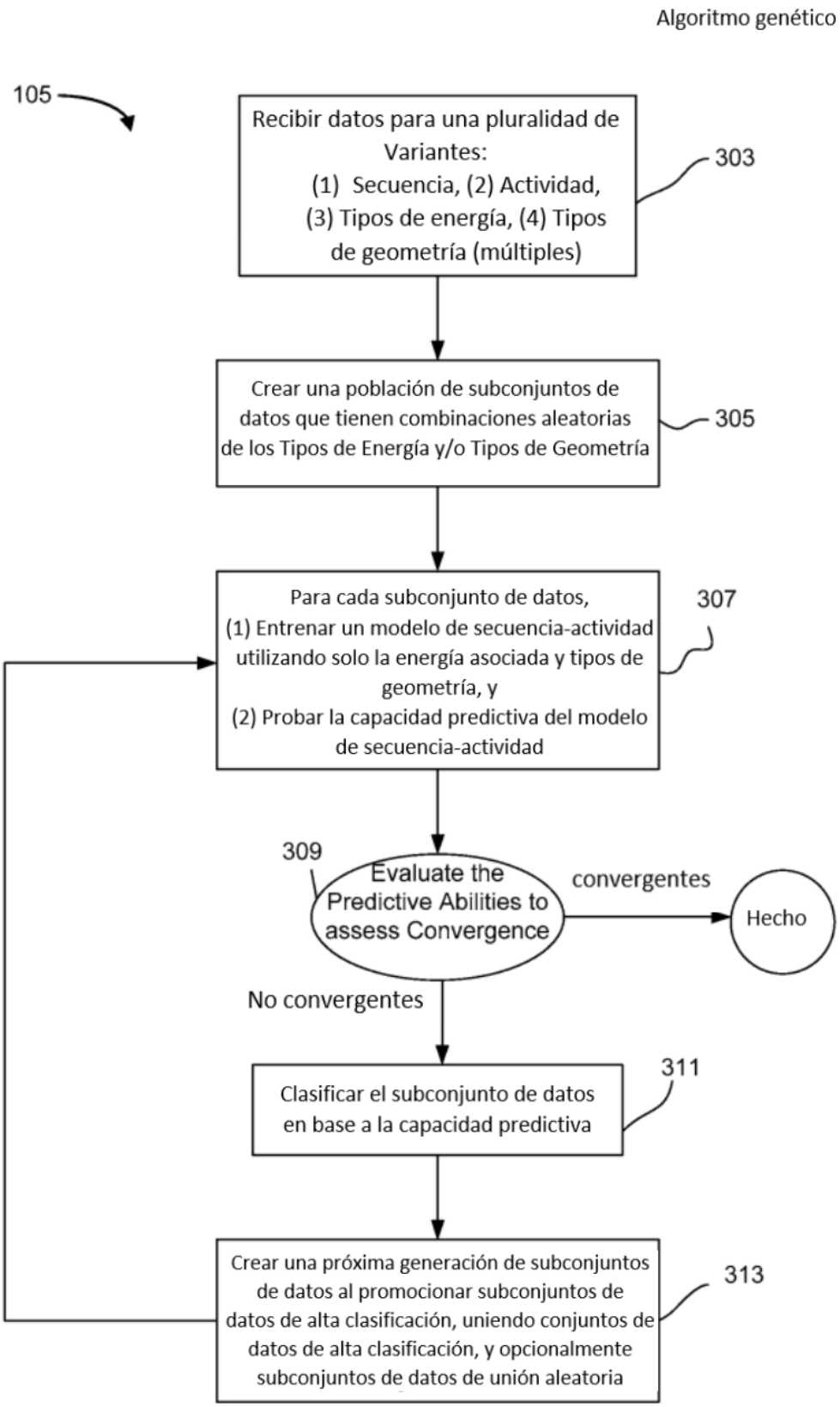
Algoritmo genético 1
105
Recibir datos para una pluralidad de Variantes:
(1) Secuencia, (2) Actividad,
(3) Tipos de energía, (4) Tipos de geometría (múltiples)
303
Crear una población de subconjuntos de datos que tienen combinaciones aleatorias de los Tipos de Energía y/o Tipos de Geometría
305
Para cada subconjunto de datos,
(1) Entrenar un modelo de secuencia-actividad utilizando solo la energía asociada y tipos de geometría, y
(2) Probar la capacidad predictiva del modelo de secuencia-actividad
307
309
Evaluate the Predictive Abilities to assess Convergence
convergentes
Hecho
No convergentes
Clasificar el subconjunto de datos en base a la capacidad predictiva
311
Crear una próxima generación de subconjuntos de datos al promocionar subconjuntos de datos de alta clasificación, uniendo conjuntos de datos de alta clasificación, y opcionalmente subconjuntos de datos de unión aleatoria
313

FIG. 3A

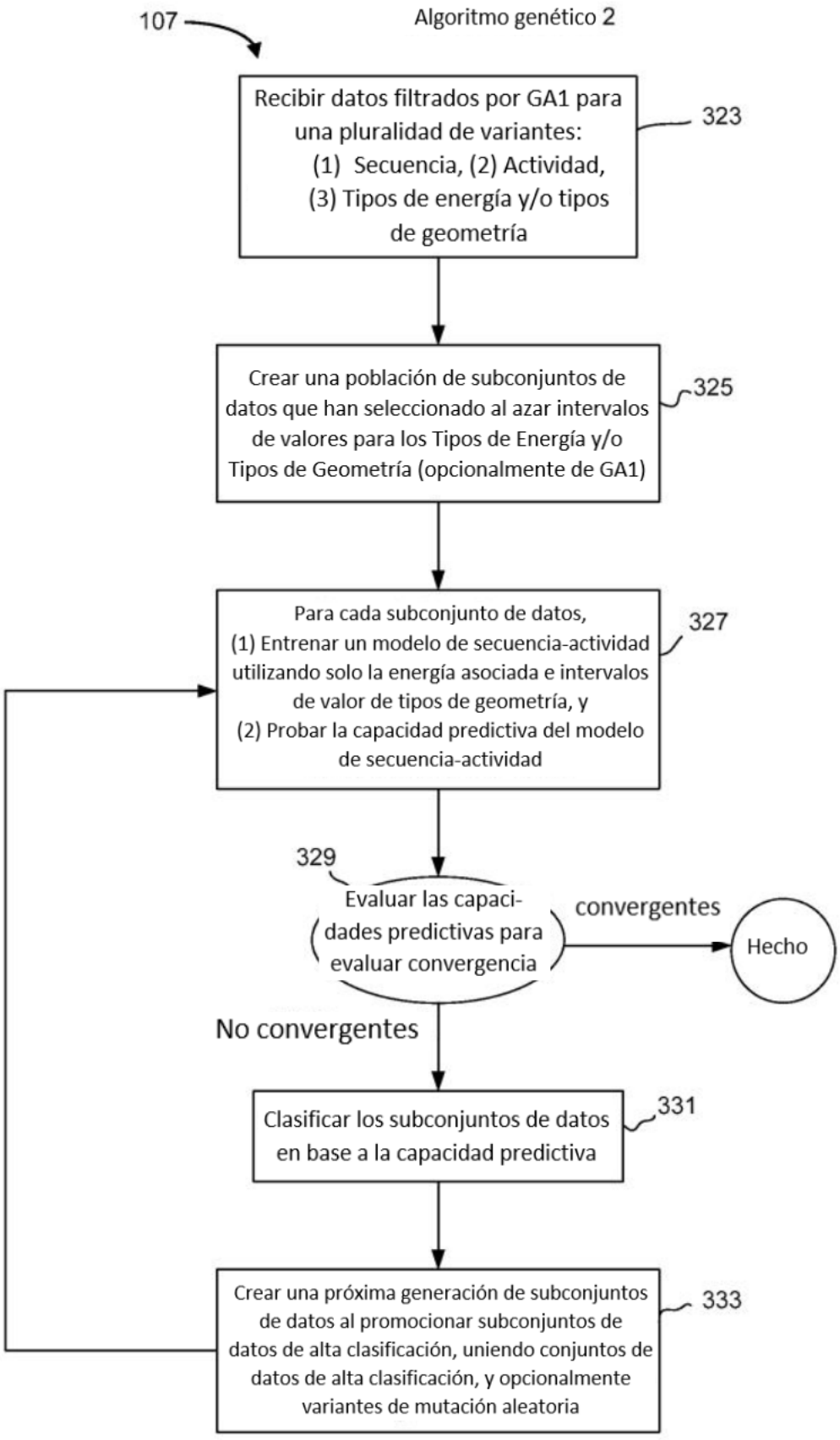
107
Algoritmo genético 2
Recibir datos filtrados por GA1 para una pluralidad de variantes: (1) Secuencia, (2) Actividad, (3) Tipos de energía y/o tipos de geometría
323
Crear una población de subconjuntos de datos que han seleccionado al azar intervalos de valores para los Tipos de Energía y/o Tipos de Geometría (opcionalmente de GA1)
325
Para cada subconjunto de datos, (1) Entrenar un modelo de secuencia-actividad utilizando solo la energía asociada e intervalos de valor de tipos de geometría, y (2) Probar la capacidad predictiva del modelo de secuencia-actividad
327
329
Evaluar las capacidades predictivas para evaluar convergencia
convergentes
Hecho
No convergentes
Clasificar los subconjuntos de datos en base a la capacidad predictiva
331
Crear una próxima generación de subconjuntos de datos al promocionar subconjuntos de datos de alta clasificación, uniendo conjuntos de datos de alta clasificación, y opcionalmente variantes de mutación aleatoria
333

FIG. 3B

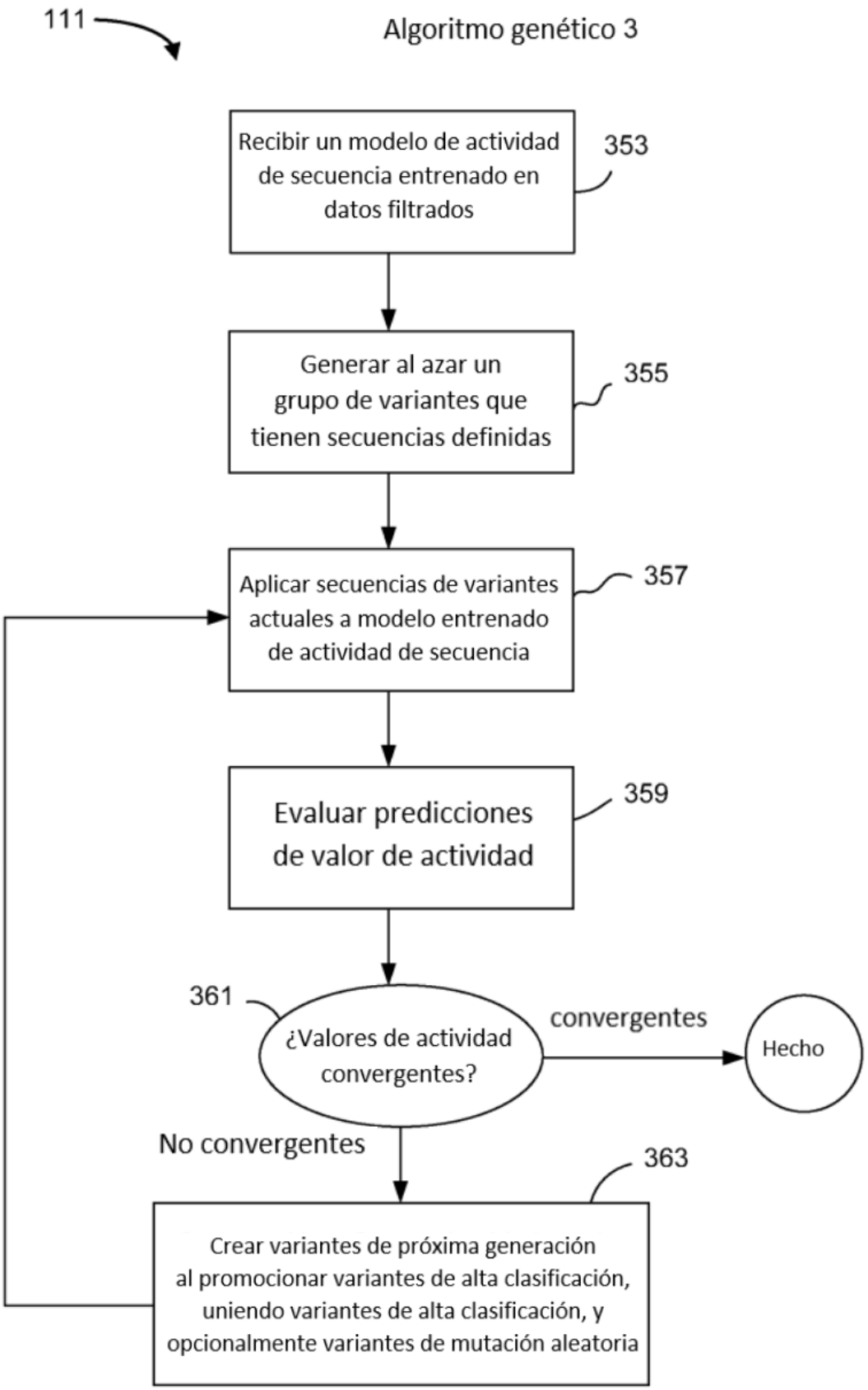
111
Algoritmo genético 3
Recibir un modelo de actividad de secuencia entrenado en datos filtrados
353
Generar al azar un grupo de variantes que tienen secuencias definidas
355
Aplicar secuencias de variantes actuales a modelo entrenado de actividad de secuencia
357
Evaluar predicciones de valor de actividad
359
361
¿Valores de actividad convergentes?
convergentes
Hecho
No convergentes
363
Crear variantes de próxima generación al promocionar variantes de alta clasificación, uniendo variantes de alta clasificación, y opcionalmente variantes de mutación aleatoria

FIG. 3C

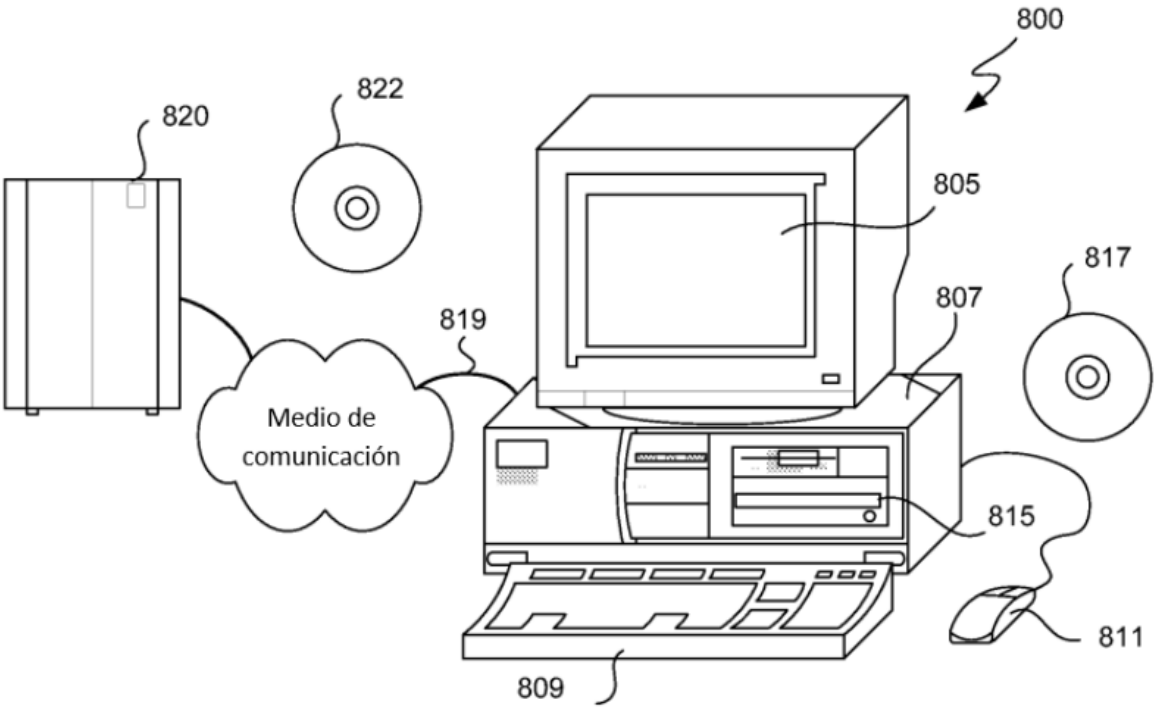

FIG. 4