(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 693 150**

(51) Int. CI.:

G06F 19/00 (2008.01)
G06F 19/16 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **26.09.2014 PCT/US2014/057899**

(87) Fecha y número de publicación internacional: **02.04.2015 WO15048572**

(96) Fecha de presentación y número de la solicitud europea: **26.09.2014 E 14781426 (3)**

(97) Fecha y número de publicación de la concesión europea: **08.08.2018 EP 3049973**

(54) Título: **Filtración automática de variantes de enzimas**

(30) Prioridad:

27.09.2013 US 201361883838 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.12.2018

(73) Titular/es:

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

(72) Inventor/es:

ZHANG, XIYUN;
SARMIENTO, RUSSELL JAVINIAR;
BASKERVILLE, DONALD SCOTT y
HUISMAN, GJAIT W.

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Filtración automática de variantes de enzimas

FONDO

[0001] Diseño de proteínas durante mucho tiempo ha sido conocido por ser una tarea difícil, si no por otra razón que la explosión combinatoria de posibles moléculas que constituyen el espacio de secuencias de búsqueda. El espacio de secuencia de las proteínas es inmenso y es imposible explorar exhaustivamente usando 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 del gran número de variantes polipeptídicas que deben ser secuenciadas, cribadas y analizadas. Los métodos de evolución dirigidos aumentan la eficacia en el afilado de las biomoléculas candidatas que tienen propiedades ventajosas. En la actualidad, la evolución dirigida de las proteínas está dominada por varios formatos de selección y recombinación de alto rendimiento, a menudo de forma iterativa.

[0002] También se han propuesto diversas técnicas computacionales para explorar el espacio de secuencias-actividad. Hablando relativamente, estas técnicas están en su infancia y todavía se necesitan avances significativos. De acuerdo con esto, son muy deseables nuevos métodos para mejorar la eficacia de exploración, secuenciación y ensayo de biomoléculas candidatas.

[0003] Hediger y col. (PLoS One. 2012; 7 (12): e49849.) Describe un método computacional para evaluar la actividad enzimática, que se basa en los métodos PM6 y MOZYME implementados en MOPAC2009, y se prueba en el primer paso de la reacción de hidrólisis de amida catalizada por la enzima Candida Antarctica lipasa B (CalB).

RESUMEN

[0004] La invención proporciona un método, implementado usando un sistema informático que incluye uno o más procesadores y la memoria del sistema, para el cribado de una pluralidad de diferentes enzimas variantes para la actividad con un sustrato, en el que la pluralidad de diferentes variantes de la enzima comprende al menos diez diferentes variantes, y las variantes enzimáticas comprenden sitios activos que difieren entre sí por al menos una mutación en la secuencia de aminoácidos del sitio activo, comprendiendo el método:

(a) crear o recibir un modelo estructural para cada una de la pluralidad de diferentes variantes de enzima, en el que cada modelo estructural contiene una representación computacional tridimensional de un sitio activo de una variante de enzima;

(b) para cada variante enzimática, acoplamiento, por el sistema informático, una representación computacional del sustrato a la representación computacional tridimensional del sitio activo de la variante enzimática, donde el acoplamiento (i) genera una pluralidad de posiciones del sustrato en el sitio activo, en donde una posición comprende una posición u orientación del sustrato con respecto al sitio activo de la variante enzimática, y (ii) identifica posiciones energéticamente favorables del sustrato en el sitio activo, en donde una posición energéticamente favorable es una posición que tiene una energía que es favorable para la unión entre el sustrato y la variante enzimática;

(c) para cada posición energéticamente favorable, determinar si la posición está activa, en donde una posición activa cumple una o más restricciones para que el sustrato experimente una reacción catalítica en el sitio activo; y

(d) seleccionar al menos una de las variantes de enzima que tiene un sitio activo en el que el sustrato tiene una o más posiciones activas como se determina en (c).

[0005] La invención proporciona además un producto de programa de ordenador que comprende uno o más medios de almacenamiento transitorio no legible por ordenador que tiene almacenado en el mismo instrucciones ejecutables por ordenador que, cuando son ejecutadas por uno o más procesadores de un sistema informático, dan lugar a que el sistema de ordenador implemente el método del modo descrito arriba.

[0006] La invención proporciona además un sistema, que comprende:

uno o más procesadores;

memoria del sistema; y

en donde uno o más procesadores y memoria están configurados para implementar un método como se indicó anteriormente.

[0007] La presente descripción se refiere a los campos de la biología molecular, evolución molecular, bioinformática

y sistemas digitales. También se proporcionan sistemas, incluidos sistemas digitales, y software de sistema para realizar estos métodos. 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 actividad y selectividad deseadas para reacciones catalíticas de sustratos particulares.

[0008] Ciertos aspectos de la presente descripción se refieren a métodos para proteínas prácticamente de detección que tienen propiedades beneficiosas y/o de guía de 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 de biomoléculas complejas o conjuntos de tales bibliotecas. Algunas realizaciones de la presente divulgación proporcionan métodos para seleccionar virtualmente enzimas para la actividad y selectividad deseadas para reacciones catalíticas en sustratos particulares. Algunas realizaciones combinan exploración y evolución dirigida para diseñar y desarrollar proteínas y enzimas que tienen propiedades deseadas. También se proporcionan productos de sistemas y programas informáticos que implementan los métodos.

[0009] Algunas realizaciones de la descripción proporcionan métodos para el cribado de una pluralidad de diferentes variantes de la enzima para la actividad con un sustrato. 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) para cada variante de enzima, acoplamiento, por el sistema informático, una representación computacional del sustrato a una representación computacional de un sitio activo de la variante enzimática, en donde el acoplamiento (i) genera una pluralidad de posiciones del sustrato en el sitio activo, e (ii) identifica posiciones energéticamente favorables del sustrato en el sitio activo; (b) para cada posición energéticamente favorable, determinar si la posición es activa, en donde una posición activa cumple una o más restricciones para que el sustrato experimente catálisis en el sitio activo; y (c) selecciona al menos una de las variantes de enzima que se determina que tiene una o más posiciones activas.

[0010] En algunas realizaciones, las limitaciones incluyen uno o más de los siguientes: restricciones de posición, distancia, ángulo, y de torsión. En algunas realizaciones, las restricciones incluyen una distancia entre un resto particular en el sustrato y un resto particular o resto en el sitio activo. En algunas realizaciones, las restricciones incluyen una distancia entre un resto particular en el ligando y un ligando nativo posicionado idealmente en el sitio activo.

[0011] En algunas realizaciones, la representación computacional del sustrato representa una especie a lo largo de la coordenada de reacción para la actividad de la enzima. La especie se selecciona del sustrato, un intermedio de reacción del sustrato, o un estado de transición del sustrato. En algunas realizaciones, las variantes rastreadas se seleccionan de un panel de enzimas que puede convertir múltiples sustratos y en donde los miembros del panel poseen al menos una mutación con respecto a una secuencia de referencia. En algunas realizaciones, al menos una mutación es una mutación de único residuo. En algunas realizaciones, al menos una mutación está en el sitio activo de la enzima. En algunas realizaciones, la pluralidad de variantes incluye una o más enzimas que pueden catalizar una reacción química seleccionada entre reducción de cetona, transaminación, oxidación, hidrólisis de nitrilo, reducción de imina, reducción de enona, hidrólisis de acilo y deshalogenación de halohidrina. En algunas realizaciones, la enzima se selecciona de reductasa de cetona, transaminasa, citocromo P450, monooxigenasa de Baeyer-Villiger, monoaminooxidasa, nitrilasa, reductasa de imina, reductasa de enona, acilasa y halohidrina deshalogenasa. Sin embargo, no se pretende que la presente invención se limite a ninguna enzima o clase de enzima particular, ya que cualquier enzima adecuada encuentra uso en los métodos de la presente invención. En algunas realizaciones, las variantes son miembros de la biblioteca producida por una o más rondas de evolución dirigida in vitro y/o in silico.

[0012] En algunas realizaciones, el método filtra al menos unos diez variantes diferentes. En otras realizaciones, el método selecciona al menos aproximadamente mil variantes diferentes.

[0013] En algunas realizaciones, las representaciones computacionales de sitios activos están dentro de modelos de homología 3-D para la pluralidad de variantes. En algunas realizaciones, se proporcionan métodos para producir los modelos de homología 3-D para variantes de proteínas. En algunas realizaciones, el método se aplica para seleccionar una pluralidad de sustratos.

[0014] Algunas realizaciones proporcionan un método para identificar las limitaciones para el sustrato a someterse a la transformación química catalizada mediante la identificación de una o más posiciones de un sustrato nativo, un intermedio de reacción del sustrato nativo, o un estado de transición del sustrato nativo cuando el nativo sustrato sufre la transformación química catalizada por una enzima de tipo salvaje.

[0015] Algunas realizaciones proporcionan un método para aplicar un conjunto de una o más restricciones de la enzima a la pluralidad de variantes de la enzima, en donde las una o más restricciones de enzimas son similares a las limitaciones de una enzima de tipo salvaje cuando un sustrato nativo sufre una transformación química catalizada en presencia de la enzima de tipo salvaje.

[0016] En algunas realizaciones, la pluralidad de posturas del sustrato se obtiene mediante operaciones de conexión

que incluyen uno o más de los siguientes: red basada en la dinámica molecular de alta temperatura, rotación aleatoria, el refinamiento por recocido simulado basado en la red, y una minimización de campo final o de fuerza completa. En algunas realizaciones, la pluralidad de posiciones del ligando comprende al menos aproximadamente 10 posiciones del sustrato en el sitio activo.

[0017] En algunas realizaciones, la selección de variantes en (c) anterior implica la identificación de variantes que se ha determinado que tienen un gran número de posiciones activas por comparación con otras variantes. En algunas realizaciones, la selección en (c) implica clasificar las variantes por uno o más de los siguientes: el número de posiciones activas que tienen las variantes, puntuaciones de atraque de las posiciones activas y energías de enlace de las posiciones activas. Entonces las variantes se seleccionan según el rango. En algunas realizaciones, los puntajes de atraque se basan en la fuerza de van de Waals y la interacción electrostática. En algunas realizaciones, las energías de enlace se basan en uno o más de las siguientes: fuerza de van der Waals, interacción electrostática y energía de solvatación.

[0018] En algunas realizaciones, el método de selección también implica preparar una pluralidad de oligonucleótidos que contienen o que codifican al menos una porción de al menos una variante seleccionada. El método además implica realizar una o más rondas de evolución dirigida utilizando la pluralidad de oligonucleótidos. En algunas realizaciones, la preparación de una pluralidad de oligonucleótidos implica sintetizar los oligonucleótidos usando un sintetizador de ácido nucleico. En algunas realizaciones, realizar una o más rondas de evolución dirigida comprende fragmentar y recombinar la pluralidad de oligonucleótidos. En algunas realizaciones, realizar una o más rondas de evolución dirigida implica realizar mutagénesis de saturación en la pluralidad de oligonucleótidos.

[0019] En algunas realizaciones, la variante de la enzima filtrada ha deseado actividad catalítica y/o selectividad. El método de algunas formas de realización también implica sintetizar la enzima seleccionada a partir del cribado.

[0020] En algunas realizaciones, el método de cribado se puede ampliar para detectar biomoléculas distintas de enzimas. Algunas realizaciones proporcionan un método para examinar una pluralidad de variantes de proteínas para la interacción con un ligando. El método implica: (a) para cada variante de proteína, acoplamiento, por el sistema informático, una representación computacional del ligando a una representación computacional de un sitio activo de la variante enzimática, en donde el acoplamiento (i) genera una pluralidad de posiciones del ligando en el sitio activo, e (ii) identifica posiciones energéticamente favorables del ligando en el sitio activo; (b) para cada posición energéticamente favorable, determinar si la posición es activa, en donde una posición activa cumple una o más restricciones para que el ligando experimente una interacción particular con la variante de proteína; y (c) seleccionar al menos una de las variantes de proteína que se determina que tiene una o más posiciones activas. En algunas realizaciones, el ligando puede seleccionarse de un sustrato, un intermedio, un estado de transición, un producto, un inhibidor, un agonista y/o un antagonista.

[0021] En algunas realizaciones, también se proporcionan productos de programas de ordenador y sistemas informáticos de aplicación de los métodos para el cribado de enzimas y proteínas.

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

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0023]

La Figura 1 ilustra las restricciones geométricas para identificar las posiciones activas para una reacción catalítica de selectividad pro-R, la reacción que implica una enzima cetona-reductasa con un resto de tirosina, un sustrato de acetofenona y el cofactor NADPH.
La Figura 2 es un diagrama de flujo que presenta un flujo de trabajo para analizar la actividad potencial de biomoléculas candidatas en algunas implementaciones.
La Figura 3A es un diagrama de flujo que muestra un ejemplo de un flujo de trabajo para diseñar secuencias de biomoléculas de acuerdo con algunas realizaciones de la divulgación.
La Figura 3B es un diagrama de flujo que muestra un ejemplo de un flujo de trabajo para diseñar secuencias de biomoléculas, que implica sintetizar y analizar secuencias obtenidas a partir de cribado virtual.
La Figura 3C es un diagrama de flujo que muestra un ejemplo de un flujo de trabajo para diseñar secuencias de biomoléculas, que combina la evolución dirigida in vitro y el cribado virtual en cada ronda de iteraciones múltiples.
La Figura 4 muestra un dispositivo digital ejemplar que puede implementarse de acuerdo con algunas realizaciones de la presente divulgación.
La Figura 5 proporciona un gráfico de datos que muestra la energía de enlace y la selectividad de 10 mejores variantes de una segunda ronda de evolución dirigida y las cadenas principales para la ronda 1 (Rd1BB) y la ronda 2 (Rd2BB).
La Figura 6A muestra la aptitud del modelo de un modelo de actividad de secuencia construido usando datos de un sistema de exploración de proteína virtual de acuerdo con algunas realizaciones.
La Figura 6B muestra datos de validación cruzada que indican que el modelo de actividad de secuencia tal como

se construyó en la Figura 6A fue preciso para predecir la energía de unión.
La Figura 6C muestra los coeficientes para diversas mutaciones de acuerdo con el modelo de actividad de secuencia tal como se construye en la Figura 6A.
La Figura 7 muestra las cantidades que indican la conversión en el eje X y la selectividad en el eje Y desde virtualmente el cribado de las variantes de la cetoductasa para la producción enantioseletiva de (*R*)-1,1,1-trifluropropano-2-ol a partir de 1,1,1-trifluropropano-2-ona.
La Figura 8 muestra las cantidades que indican conversión y éxitos (variantes con cierto nivel de mejora) de la evolución virtual dirigida de P450 para la oxidación de CH regioselectiva a C-OH.

DESCRIPCIÓN DETALLADA

[0024] El cribado de proteínas y enzimas se puede realizar de maneras reales que implican mediciones de las propiedades químicas y físicas de proteína y moléculas de enzima que interactúan con ligandos y sustratos. Las medidas reales consumen tiempo y recursos, y los mecanismos físicos y químicos subyacentes a menudo son difíciles de visualizar o manipular. Los métodos y sistemas de cribado "virtuales" descritos en este documento proporcionan herramientas para visualizar o manipular la estructura y la dinámica de enzimas, proteínas y sus sustratos y ligandos. Estas herramientas pueden ahorrar tiempo y/o materiales para estudiar las moléculas.

[0025] En algunas realizaciones, el cribado virtual de proteínas o enzimas se utiliza en la evolución dirigida de proteínas de interés. El cribado virtual se usa en lugar del cribado físico durante varias etapas de estas formas de evolución dirigidas, lo que permite estudiar un gran número de moléculas y reacciones sin requerir los materiales físicos o el tiempo requerido por el cribado real. Estas realizaciones pueden acelerar los procesos para obtener proteínas y enzimas que tienen propiedades deseadas. Los materiales y recursos también se pueden guardar en los procesos. 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

[0026] 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. Diversos diccionarios científicos que incluyen los términos incluidos en este documento son bien conocidos y están disponibles para los expertos en la técnica. Cualquier método y materiales similares o equivalentes a los descritos en este documento encuentran uso en la práctica de las realizaciones descritas aquí.

[0027] Los términos definidos inmediatamente a continuación están más completamente entendidos por referencia a la especificación como un todo. Las definiciones tienen el propósito de describir realizaciones particulares solamente y ayudar a comprender los conceptos complejos descritos en esta memoria descriptiva. No están destinados a limitar el alcance completo de la divulgación. Específicamente, debe entenderse que esta divulgación no está limitada a las secuencias, composiciones, algoritmos, sistemas, metodología, protocolos y reactivos particulares descritos, ya que pueden variar, dependiendo del contexto en el que los utilizan los expertos en la técnica.

[0028] 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 indiquen claramente lo contrario. Por lo tanto, 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 ser utilizada en su sentido correcto como operador lógico booleano, abarcando tanto la selección de características en la alternativa (A o B, donde la selección de A es mutuamente excluyente de B) como la selección de características en conjunción (A o B, donde se seleccionan A y B).

[0029] "Acoplamiento" como se usa en el presente documento, se refiere al proceso de cálculo para simular y/o caracterizar la unión de una representación computacional de una molécula (por ejemplo, un sustrato o ligando) a una representación computacional de un sitio activo de una biomolécula (por ejemplo, una enzima o proteína). El acoplamiento se implementa típicamente en un sistema informático que utiliza un programa informático "acoplador". Típicamente, el resultado de un proceso de acoplamiento es una representación computacional de la molécula "acoplada" en el sitio activo en una "posición" específica. Se puede llevar a cabo una pluralidad de procesos de acoplamiento entre la misma representación computacional de una molécula y la misma representación computacional de un sitio activo, dando como resultado una pluralidad de "posiciones" diferentes de la molécula en el sitio activo. La evaluación de la estructura, conformación y energía de la pluralidad de diferentes "posiciones" en la representación computacional del sitio activo pueden identificar ciertas "posiciones" como más energéticamente favorables para la unión entre el ligando y la biomolécula.

[0030] En algunas realizaciones, posiciones generadas a partir de acoplamiento son evaluadas para determinar si son "activas" para una interacción deseada con la biomolécula. Las "posiciones activas" son aquellas que cumplen una o más restricciones para una actividad en consideración. Una "restricción" puede limitar la estructura, geometría, conformación, energía, etc. de una posición. En ciertas formas de realización, una "posición activa" de una representación computacional de un sustrato en el sitio activo de una enzima satisface las condiciones para la

catálisis por la enzima. Cuando el acoplamiento identifica numerosas posiciones activas de una representación computacional de un sustrato en la representación computacional del sitio activo, la enzima específica representada puede seleccionarse como favorable para catalizar la transformación química del sustrato al producto.

[0031] Un "acoplador" es un programa informático que computacionalmente simula y/o caracteriza el proceso de acoplamiento entre una representación computacional de una molécula (por ejemplo, un sustrato o ligando) y una representación computacional de un sitio activo de interés en una proteína u otra molécula biológica.

[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 comercialmente disponibles incluyen CDocker (Accelrys), DOCK (Universidad de California, San Francisco), AutoDock (Instituto de Investigación Scripps), FlexX (tripos.com), GOLD (ccdc.cam.ac.uk) y GLIDE (schrodinger. com).

[0033] El acoplamiento usando un acoplador típicamente genera "posiciones" de representaciones computacionales de sustratos y ligandos con respecto a los sitios activos. Estas posturas pueden usarse para generar un puntaje de atraque o para evaluar el atraque. En algunas realizaciones, las posturas están asociadas con los valores de energía de interacción calculados por un acoplador. Algunas posiciones son enérgicamente más favorables que otras posiciones. En algunas realizaciones, el acoplador permite a un usuario especificar un número de posiciones (n) para uso al evaluar el acoplamiento. Solo se consideran las posiciones n superiores con los mejores puntajes de atraque al evaluar el atraque. En algunas realizaciones, solo las posiciones con energía de interacción favorable que cumplen los criterios definidos se seleccionan para clasificarse como posiciones activas o inactivas.

[0034] En algunas realizaciones, un acoplador puede determinar que es probable que un sustrato o ligando se una con una biomolécula si una o más posiciones del sustrato o ligando tienen energía de interacción favorable con la biomolécula. Un ligando unido puede actuar como un agonista o antagonista. Varios estibadores emiten un puntaje de atraque u otra medida de unión entre el sustrato o ligando y la biomolécula. Para algunas combinaciones de sitios activos de biomoléculas con un sustrato o ligando, el programa de acoplamiento determinará que es poco probable que se produzca la unión. En tales casos, el programa de acoplamiento arrojará una conclusión de que el sustrato o ligando no se une a la biomolécula.

[0035] Un acoplador puede ser programado para emitir una evaluación de la probabilidad de que un ligando se acoplará con el sitio activo de biomolécula o la calidad de tal acoplamiento, en caso de producirse. La probabilidad y la calidad del acoplamiento indican la probabilidad de que un ligando se una con una biomolécula. En un nivel, un acoplador determina si es probable que un ligando se una al sitio activo de una biomolécula. Si la lógica de acoplamiento concluye que el enlace no es probable o es altamente desfavorable, puede generar un resultado de "no se encuentran posiciones refinadas". Esto puede ocurrir cuando todas las conformaciones generadas por el programa de acoplamiento tienen enfrentamientos de van der Waals desfavorables y/o repulsiones electrostáticas con el sitio activo. En el ejemplo anterior de un procedimiento de acoplamiento, si la segunda operación no logra encontrar una posición con energía suave inferior al umbral, el acoplador puede devolver un resultado como "no se encontraron posiciones refinadas". Debido a que la energía suave considera principalmente las interacciones no fusionadas que incluyen van der Waals y las fuerzas electrostáticas, el resultado "no se encontró una posición refinada" significa que el ligando tiene enfrentamientos estéricos severos y/o repulsiones electrostáticas con el receptor de biomoléculas para un número determinado de posiciones.

[0036] En ciertas realizaciones, el acoplador da salida a una puntuación de acoplamiento que representa la interacción entre el ligando en el sitio activo de biomolécula. Los estibadores 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 usando un campo de fuerza como CHARMm.

[0037] En diversas realizaciones, los programas de conexión generan tales salidas al considerar múltiples posiciones del ligando en el sitio activo de la biomolécula. Cada posición tendrá sus propios valores de energía asociados. En algunas realizaciones, el programa de acoplamiento clasifica las posiciones y considera la energía asociada con una o más de las posiciones de alto rango. En algunos casos, puede promediar las energías de ciertas posiciones de alto rango o realizar un análisis estadístico de las mejores posiciones de clasificación. En otras realizaciones, simplemente elige el valor asistido con la posición superiormente clasificada y lo emite como la energía resultante para el acoplamiento.

[0038] En algunas realizaciones, la representación computacional de un sustrato corresponde a una especie molecular a lo largo de la coordenada de reacción de una reacción enzimática que es capaz de convertir la molécula de sustrato a la molécula de producto deseado. En algunas realizaciones, la representación computacional del sustrato representa la molécula de sustrato per se. En algunas realizaciones, la representación computacional del sustrato representa una estructura intermedia del sustrato que se forma a lo largo de la coordenada de reacción (es decir, un "intermedio de reacción del sustrato"). En algunas realizaciones, la representación computacional del

sustrato representa una estructura de estado de transición que se forma a lo largo de la coordenada de reacción enzimática (es decir, un "estado de transición del sustrato").

[0039] En algunas realizaciones, una representación computacional de un ligando puede representar una especie molecular que se une fuertemente a una enzima o biomolécula pero no procede a lo largo de una coordenada de reacción a un producto deseado. Por ejemplo, la representación computacional del ligando puede representar un fuerte inhibidor para detectar inhibidores de una enzima, o antagonistas de unión fuerte o agonistas de proteínas (por ejemplo, receptores).

[0040] Una "posición" es la posición u orientación de un sustrato o ligando con respecto a un sitio activo de una molécula biológica. En una posición, 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 activo. Si bien la conformación de un ligando no es su posición, porque la conformación no considera el sitio activo, la conformación puede usarse para determinar una posición. En algunas realizaciones, la orientación y conformación de un ligando definen una posición. En algunas realizaciones, una posición solo existe si la combinación de orientación/conformación de un ligando cumple un nivel de energía umbral definido en el sitio activo de referencia.

[0041] Varios mecanismos computacionales se pueden emplear para generar posiciones para acoplamiento. Los ejemplos incluyen búsquedas de torsión sistemáticas o estocásticas sobre enlaces giratorios, simulaciones de dinámica molecular y algoritmos genéticos para "evolucionar" nuevas conformaciones de baja energía. Estas técnicas se utilizan para modificar representaciones computacionales del ligando y/o sitio activo para explorar el "espacio de posición".

[0042] Los acopladores evalúan las posiciones para determinar cómo interactúa el ligando con el sitio activo. En algunas realizaciones, lo hacen 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 usa para caracterizar el acoplamiento y en algunos casos produce una puntuación de atraque. En algunas implementaciones, los estibadores clasifican las posiciones basadas en puntajes de atraque. En algunas implementaciones, los estibadores eliminan las posiciones con puntuaciones de atraque desfavorables de consideración.

[0043] En ciertas realizaciones, un sistema de detección de la proteína virtual evalúa una posición para determinar si la posición está activa. Se considera que una posición está activa si cumple con las restricciones definidas que se sabe que son importantes para la actividad deseada bajo consideración. Como ejemplo, el sistema de selección de proteína virtual puede determinar si una posición soporta la transformación catalítica del ligando en un sitio activo.

[0044] Un "ligando" es una molécula o complejo que interactúa con un sitio activo 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 (algunas veces requiere) otras entidades químicas tales como cofactores orgánicos e inorgánicos (por ejemplo, coenzimas y grupos prostéticos), iones metálicos y similares. Los ligandos pueden ser agonistas o antagonistas.

[0045] El "sitio activo" de una biomolécula es un sitio definido por la estructura de la biomolécula que es capaz de contener y/o unir la totalidad o parte de una molécula (por ejemplo, un sustrato o ligando). Se contemplan muchos tipos de sitios activos y algunos de estos se describen en otra parte del presente documento. A menudo, el sitio activo contiene características químicas y/o físicas (por ejemplo, residuos de aminoácidos) capaces de formar interacciones de unión con el sustrato o ligando. En algunas realizaciones (por ejemplo, cuando la biomolécula es una enzima), el "sitio activo" incluye al menos un residuo catalítico y una pluralidad de residuos de unión, y a veces otras entidades químicas tales como cofactores orgánicos e inorgánicos (por ejemplo, coenzimas y grupos prostéticos), iones metálicos y similares. Al menos un residuo catalítico del sitio activo puede contener un resto catalítico que cataliza la renovación de un sustrato. Los residuos de unión del sitio activo proporcionan interacciones de unión con el sustrato para mantenerlo en el sitio activo de una manera estereoselectiva y/o regioselectiva. Dichas interacciones pueden incluir interacciones de van der Waals, interacciones electrostáticas, enlaces de hidrógeno, interacciones hidrofílicas, interacciones hidrofóbicas, interacciones de solventes, enlaces covalentes, etc.

[0046] En algunas realizaciones, una representación computacional de un sitio activo se puede utilizar para acoplar una representación computacional de un sustrato o ligando, generando así posiciones que pueden evaluarse para una interacción favorable con el sitio activo (p. ej., determinación de la energía de enlace para las posiciones).

[0047] En algunas realizaciones, la representación computacional del sitio activo se define geométricamente por una esfera u otra forma. En algunas realizaciones, el sitio activo se define creando una esfera alrededor del centro de objetos seleccionados (p. ej., ligandos y/u otras entidades químicas en el molde de estructura) con el radio ajustado para incluirlos. El radio mínimo es de 5Å, pero el tamaño del sitio activo se puede expandir aumentando el radio de la esfera en 1Å, 2Å, 3Å, 4Å, 6Å, 8Å, 10Å, y así sucesivamente. En algunas implementaciones, el tamaño del radio se selecciona para capturar residuos próximos al sustrato. Por lo tanto, los sustratos más grandes se asociarán con radios más grandes y los sustratos pequeños se asociarán con radios más pequeños. No se pretende que la presente divulgación se limite a ningún valor particular de radios. En algunas realizaciones, el sitio activo se puede definir a partir de cavidades de receptor, donde el sitio activo se derivó de una de las cavidades detectadas en el molde de estructura. En algunas realizaciones, el sitio activo se puede definir a partir de los registros del sitio de

Protein Data Bank (PDB), ya que el archivo PDB del molde de estructura a menudo tiene un sitio activo definido usando registros del sitio. Dado que todos los modelos de homología se crearán utilizando el molde de estructura, el sitio activo definido es transferible a todos los modelos de homología.

[0048] En algunas realizaciones, la representación computacional del sitio activo puede ser definida por varias formas tridimensionales, tales como una forma personalizable por el usuario (por ejemplo, una elipse o una forma irregular que refleja la estructura del sustrato) con referencia a restos en el sustrato y/o la enzima.

[0049] En algunas realizaciones, la representación computacional del sitio activo puede ser definida para incluir aminoácidos que no interactúan directamente (por ejemplo, a través de interacciones de van der Waals, interacciones electrostáticas, enlaces de hidrógeno) con el sustrato o molécula de ligando en el activo sitio, pero que interactúan con otros aminoácidos en la representación computacional del sitio activo, y por lo tanto afectan la evaluación de posiciones del sustrato o ligando.

[0050] En algunas realizaciones, los residuos que contribuyen a la catálisis y/o unión pueden existir fuera de la representación computacional del sitio activo como se definió anteriormente. Dichos residuos pueden modificarse durante la evolución dirigida considerando residuos más allá del sitio activo como candidatos para mutación o recombinación.

[0051] 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 colisionantes proceden a formar productos. En esta descripción, algunas veces cuando se describe un sustrato en un proceso, el estado intermedio y de transición también puede ser adecuado para el proceso. En tales situaciones, el sustrato, el intermedio y el estado de transición se pueden denominar colectivamente como "ligandos". En algunos casos, se generan intermedios múltiples en la transformación catalítica de un sustrato. En ciertas realizaciones, la especie de ligando (sustrato o estado intermedio o de transición) elegida para el análisis es una que se sabe que está asociada con una etapa limitante de velocidad en la transformación catalítica. Como ejemplo, un sustrato unido covalentemente a un cofactor de enzima puede modificarse químicamente en una etapa de limitación de velocidad. En tal caso, la especie sustrato-cofactor se usa para modelar la interacción.

[0052] Un "ligando" es una molécula capaz de unirse a una biomolécula y puede incluir moléculas "sustrato" que son capaces de unirse y de someterse, además, a una transformación química catalítica. Algunos ligandos se unen con un sitio activo pero no experimentan 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, un ligando se evalúa por su capacidad para potenciar, activar o inhibir el comportamiento natural de una biomolécula.

[0053] Una "biomolécula" o una "molécula biológica" se refiere a una molécula que se encuentra generalmente en o producida por un organismo biológico. En algunas realizaciones, las moléculas biológicas comprenden macromoléculas biológicas poliméricas que tienen múltiples subunidades (*es decir,* "biopolímeros"). Las biomoléculas típicas incluyen proteínas, enzimas y otros polipéptidos, ADN, ARN y otros polinucleótidos, y también pueden incluir moléculas que comparten algunas características estructurales con polímeros naturales tales como 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,* análogos de ARN, análogos de ADN, análogos de polipéptidos, ácidos nucleicos peptídicos (PNA), combinaciones de ARN y ADN (*por ejemplo,* quimeraplastos) o similares. No se pretende que las biomoléculas se limiten a cualquier molécula particular, ya que cualquier molécula biológica adecuada encuentra uso en la presente descripción, incluyendo pero no limitado a, por ejemplo, lípidos, hidratos de carbono, u otras moléculas orgánicas que son realizadas 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 biomoléculas que tienen sitios activos que interaccionan con un ligando para efectuar una transformación química o biológica, por ejemplo, catálisis de un sustrato, activación de biomoléculas, o inactivación de las biomoléculas, específicamente enzimas.

[0054] En algunas realizaciones, una "propiedad beneficiosa" o "actividad" es un aumento o disminución en una o más de las siguientes: velocidad catalítica (k_{cat}), la afinidad de unión al sustrato (K_M), la eficiencia catalítica (k_{cat}/K_M), especificidad del sustrato, quimioterapia, regioselectividad, estereoselectividad, estereoespecificidad, especificidad del ligando, agonismo del receptor, antagonismo del receptor, conversión de un cofactor, estabilidad del oxígeno, nivel de expresión de la proteína, solubilidad, termoactividad, termoestabilidad, actividad del pH, estabilidad del pH (por ejemplo, a pH alcalino o ácido), inhibición de glucosa y/o resistencia a inhibidores (por ejemplo, ácido acético, lectinas, ácidos tánicos y compuestos fenólicos) y proteasas. Otras actividades deseadas pueden incluir un perfil alterado en respuesta a un estímulo particular (p. ej., 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 rastreadas como se describe aquí 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 conversión porcentual de un cofactor. No se pretende que la presente divulgación se limite a ninguna propiedad beneficiosa particular y/o actividad deseada.

[0055] 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 en este documento incluye selectividad, aunque convencionalmente la selectividad a veces se ve como distinta de la actividad enzimática.

[0056] Los términos "proteína", "polipéptido" y "péptido" se usan indistintamente para referirse a un polímero de al menos dos aminoácidos unidos covalentemente mediante un enlace amida, independientemente de la longitud o modificación postraduccional (por ejemplo, glicosilación, fosforilación, lipidación, miristilación, ubiquitinación, etc.). En algunos casos, el polímero tiene al menos aproximadamente 30 residuos de aminoácidos, y habitualmente al menos aproximadamente 50 residuos de aminoácidos. Más típicamente, contienen al menos aproximadamente 100 residuos de aminoácidos. Los términos incluyen composiciones convencionalmente consideradas 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 en este documento no están restringidos a los aminoácidos genéticamente codificados. De hecho, además de los aminoácidos codificados genéticamente, los polipéptidos descritos en la presente memoria pueden estar compuestos, en su totalidad o en parte, de aminoácidos no codificados naturales 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, espacios) y/o sustituciones, en comparación con la secuencia de aminoácidos del polipéptido parental de longitud completa, mientras que aún retiene la actividad funcional (p. ej., actividad catalítica).

[0057] Una biomolécula u organismo de "tipo salvaje" (WT) es uno que tiene el fenotipo de la forma típica de una especie, ya que se produce en la naturaleza. En ocasiones, se ha aislado una biomolécula de tipo salvaje de una fuente natural. Otras veces, se deriva en el entorno de laboratorio. Habitualmente, las biomoléculas de tipo salvaje se relacionan o codifican por secuencias genéticas de genomas normales o de referencia en oposición a los genomas mutantes. Se incluyen dentro de la definición de "biomoléculas de tipo silvestre" las formas recombinantes de un polipéptido o polinucleótido que tiene una secuencia idéntica a la forma nativa. Un sustrato o ligando que reacciona con una biomolécula de tipo salvaje a veces se considera un sustrato o ligando "nativo".

[0058] Como se usa en el presente documento, los términos "variantes", "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 se puede referir como una "mutación". En algunas realizaciones, un mutante es un polipéptido o secuencia de polinucleótido que se ha alterado mediante al menos una sustitución, inserción, cruzamiento, deleción y/u otra operación genética. Para los fines de la presente descripción, los mutantes y las variantes no están limitadas a un método particular mediante 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 restos de aminoácidos que se han mutado, 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 invariantes, o no están mutados en comparación con un polipéptido original en los polipéptidos variantes que constituyen 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.

[0059] Como se usa en el presente documento, los términos "variante de la enzima" y "enzima 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 silvestre u otra enzima de referencia. Las variantes de enzimas pueden prepararse mediante una amplia variedad de técnicas de mutagénesis diferentes bien conocidas por los expertos en la técnica. Además, los kits de mutagénesis también están disponibles en muchos proveedores comerciales de biología molecular. Se encuentran disponibles métodos para realizar sustituciones específicas en aminoácidos definidos (dirigidos a sitio), mutaciones específicas o aleatorias en una región localizada del gen (regio-específico) o mutagénesis aleatoria sobre el gen completo (por ejemplo, mutagénesis de saturación). Numerosos métodos adecuados son conocidos por expertos en la técnica para generar variantes de enzimas, que incluyen, pero no se limitan a mutagénesis dirigida al sitio, ADN monocatenario o ADN bicatenario usando PCR, mutagénesis en casete, síntesis génica, PCR propensa a error, barajado y mutagénesis de saturación química, o cualquier otro método adecuado conocido en la técnica. Después de que se producen las variantes, se pueden seleccionar para la propiedad deseada (por ejemplo, alta o aumentada, o baja o reducida actividad, mayor estabilidad térmica y/o alcalina, etc.).

[0060] 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 girar colectivamente sobre sustratos múltiples, experimentando cada uno la misma reacción. A menudo, los miembros del panel son elegidos para entregar de manera eficiente múltiples sustratos. En algunos casos, los paneles están disponibles

comercialmente. En otros casos, son propiedad de una entidad. Por ejemplo, un panel puede incluir diversas enzimas identificadas como éxitos en un procedimiento de selección. 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.

[0061] 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 que se representa. Los modelos a menudo contienen representaciones producidas computacionalmente de los sitios activos u otros aspectos de las variantes de la enzima. Los ejemplos de modelos relevantes para las realizaciones de la presente invención se producen a partir del modelado de homología, el enhebramiento de proteínas o el modelado de proteínas *ab initio* usando una rutina tal como Rosetta (rosettacommons.org/software/) o simulaciones de Dinámica Molecular.

[0062] Un "modelo de homología" es un modelo tridimensional de una proteína o parte de una proteína que contiene al menos el sitio activo de un ligando bajo consideración. El modelo de homología se basa en la observación de que las estructuras de proteínas tienden a conservarse entre proteínas homólogas. Un modelo de homología proporciona posiciones tridimensionales de los residuos que incluyen la cadena principal y las cadenas laterales. El modelo se genera a partir de un molde de estructura de una proteína homóloga que se asemeja a la estructura de la secuencia modelada. En algunas realizaciones, se usa un molde de estructura en dos pasos: "alinear secuencia con moldes" y "construir modelos de homología".

[0063] El paso "alinear secuencia a los moldes" alinea la secuencia modelo a 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 las lagunas y otras regiones de disimilitud entre la secuencia del modelo y la(s) secuencia(s) del molde de la estructura.

[0064] El paso de "modelos de homología de construcción" utiliza características estructurales del molde de estructura para derivar las restricciones espaciales que, a su vez, se utilizan para generar, por ejemplo, estructuras de proteínas modelo utilizando gradiente conjugado y procedimientos de optimización de recocido simulado. Las características estructurales del molde se pueden obtener a partir de una técnica como la RMN o la cristalografía de rayos X. Se pueden encontrar ejemplos de tales técnicas en el artículo de revisión, "A Guide to Template Based Structure Prediction", por Qu X, Swanson R, Day R, Tsai J. Curr Protein Pept Sci. 2009 Jun; 10 (3): 270-85.

[0065] El término "conformación activa" se usa en referencia a una conformación de una proteína (por ejemplo, una enzima) que permite que la proteína de lugar a que un sustrato se someta a una transformación química (por ejemplo, una reacción catalítica).

[0066] Una "posición activa" es una en la que es probable que un ligando sufra una transformación catalítica o realice alguna función deseada tal como la unión de forma covalente con el sitio de unión de un ligando.

[0067] Los términos "oxidorreducción", "oxidación-reducción", y "redox" se usan indistintamente con referencia a una reacción química reversible en la que una reacción es una oxidación y la inversa es una reducción. Los términos también se usan para referirse a todas las reacciones químicas en las que los átomos han cambiado su estado de oxidación; en general, las reacciones redox implican la transferencia de electrones entre especies. Esto puede ser un proceso simple redox, tales como la oxidación de carbono para producir dióxido de carbono (CO_2) o la reducción de carbono con hidrógeno para producir metano (CH4), o un proceso complejo tal como la oxidación de la glucosa ($C_6H_{12}O_6$) en el cuerpo humano a través de una serie de procesos complejos de transferencia de electrones.

[0068] Una "oxidorreductasa" es una enzima que cataliza una reacción de oxidorreducción.

[0069] El término "transferación" se utiliza aquí para referirse a una reacción química que transfiere un grupo funcional a partir de un compuesto a otro compuesto. Una "transferasa" se usa para referirse a cualquiera de las diversas enzimas que catalizan una reacción de transfección.

[0070] El término "hidrólisis" se utiliza para referirse a una reacción química en la que el agua reacciona con un compuesto para producir otros compuestos, cuya reacción implica la división de un enlace químico mediante la adición del catión de hidrógeno y el anión hidróxido del agua.

[0071] Una "hidrolasa" es una enzima que cataliza una reacción de hidrólisis.

[0072] El término "isomerización" se utiliza para referirse a una reacción química que convierte un compuesto en un isómero.

[0073] Una "isomerasa" es una enzima que cataliza una reacción de isomerización, haciendo que su sustrato se convierta en una forma isomérica.

[0074] El término "ligación" se utiliza aquí para referirse a cualquier reacción química que unen dos moléculas mediante la formación de un nuevo enlace químico. En algunas realizaciones, una reacción de ligación implica la hidrólisis de un pequeño grupo químico dependiente de una de las moléculas más grandes. En algunas realizaciones, una enzima cataliza la unión de dos compuestos, por ejemplo, enzimas que catalizan la unión de CO, CS, CN, etc. Una enzima que cataliza una reacción de ligación se denomina "ligasa".

[0075] Una "liasa" es una enzima que cataliza la rotura de diversos enlaces químicos por medios distintos de la hidrólisis y oxidación. En algunas realizaciones, una reacción de liasa forma un nuevo doble enlace o una nueva estructura de anillo.

[0076] Una "cetorreductasa" es una enzima que normalmente utiliza el cofactor NADPH para reducir estereoespecíficamente un grupo ceto a un grupo hidroxilo (Véase por ejemplo, las variantes descritas en WO2008103248A2, WO2009029554A2, WO2009036404A2, WO2009042984A1, WO2009046153A1, y WO2010025238A2).

[0077] Una "transaminasa" o una "aminotransferasa" es una enzima que cataliza una reacción de transaminación entre un aminoácido y un α-cetoácido, en el que el grupo amino NH_2 en el aminoácido se intercambia con el grupo ceto =O en el α-cetoácido (véanse, por ejemplo, las variantes descritas en WO2010081053A2 y WO2010099501A2).

[0078] Las proteínas "citocromo" (abreviadas como "CYP") son enzimas involucradas en la oxidación de sustancias orgánicas. Un ejemplo son las enzimas del citocromo P450. Los sustratos de las enzimas CYP incluyen, pero no se limitan a intermedios metabólicos tales como lípidos y hormonas esteroides, así como sustancias xenobióticas tales como fármacos y otros productos químicos tóxicos. Los CYP son las principales enzimas involucradas en el metabolismo y la bioactivación de los medicamentos. Los CYP usan una variedad de moléculas pequeñas y grandes como sustratos en reacciones enzimáticas. La reacción más común catalizada por el citocromo P450 es una reacción de monooxigenasa, por ejemplo, la inserción de un átomo de oxígeno en un sustrato orgánico (RH) mientras que el otro átomo de oxígeno se reduce a agua. Las enzimas del citocromo P450 pertenecen a una superfamilia de proteínas que contienen un cofactor hemo y, por lo tanto, son hemoproteínas. En general, son enzimas oxidasa terminales en las cadenas de transferencia de electrones. Las placas y enzimas de cribado Micro-Cyp® disponibles en Codexis son útiles en la producción de metabolitos de fármacos y nuevos compuestos principales (véanse, por ejemplo, las variantes descritas en los documentos WO2002083868A2, WO2005017105A2, WO2005017116A2 y WO2003008563A2).

[0079] Una "monooxigenasa Baeyer-Villiger" es una enzima que emplea NADPH y oxígeno molecular para catalizar una reacción de oxidación de Baeyer-Villiger, en donde se inserta un átomo de oxígeno en un enlace carbono-carbono de un sustrato carbonílico (Véase por ejemplo, las variantes en WO2011071982A2 y WO2012078800A2).

[0080] Una "oxidasa de monoamina" (MAO) (EC 1.4.3.4) es una enzima que cataliza la oxidación de monoaminas, que son neurotransmisores y neuromoduladores que contienen un grupo amino que está conectado a un anillo aromático por una cadena de dos carbonos ($-CH_2-CH_2-$). Los MAO pertenecen a la familia de proteínas de oxidorreductasas de aminas que contienen flavina (véanse, por ejemplo, las variantes en WO2010008828A2).

[0081] Una "nitrilasa" o aminohidrolasa de nitrilo (EC 3.5.5.1) es una enzima que cataliza la hidrólisis de nitrilos a los ácidos carboxílicos y amoníaco, sin la formación de productos intermedios de amida "libre" (véanse, por ejemplo, las variantes en WO2011011630A2).

[0082] Una "reductasa de imina" es una enzima que cataliza la reducción de un grupo funcional de imina que contiene un doble enlace nitrógeno-carbono, rompiendo el enlace doble al dar lugar a que un electrón se done al átomo de nitrógeno.

[0083] Una "reductasa de enona" es una enzima que cataliza la reducción de un grupo funcional de enona, que incluye un sistema conjugado de un alqueno y una cetona, rompiendo el ceto o un doble enlace alqueno (Véase por ejemplo, las variantes describen en WO2010075574A2).

[0084] Una "acilasa" es una enzima que cataliza la escisión hidrolítica de amida de acilo o enlaces de éster de acilo (Véase por ejemplo, las variantes de acilasa de penicilina G en WO2010054319A2).

[0085] Una "deshalogenasa halohidrina" "HHDH" es una enzima implicada en la degradación de halohidrinas vecinales. En *Agrobacterium radiobacter* AD1, por ejemplo, cataliza la deshalogenación de halohidrinas para producir los epóxidos correspondientes (véanse, por ejemplo, las variantes descritas en WO2010080635A2).

[0086] 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, la colección de secuencias de genes para genes que interactúan, gen, secuencia de ácido nucleico, proteína, péptido, polipéptido, polisacárido, etc. En algunos contextos, una "secuencia" se refiere al orden y la identidad de los residuos de aminoácidos en una proteína (es decir, una secuencia de proteína o cadena de caracteres de proteína) o al orden

y la identidad de nucleótidos en un ácido nucleico (es decir, una secuencia de ácido nucleico o cadena de caracteres de ácido nucleico). Una secuencia puede ser 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 y la identidad de los aminoácidos que comprenden una proteína o péptido.

[0087] "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.

[0088] 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 reguladoras 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 a partir de una variedad de fuentes, incluida la clonación a partir de una fuente de interés o la síntesis a partir de información de secuencia conocida o predicha, y pueden incluir secuencias diseñadas para tener los parámetros deseados.

[0089] 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.

[0090] "Filtración" se refiere al proceso en el que se determinan una o más propiedades de una o más biomoléculas. Por ejemplo, los procesos de selección típicos incluyen aquellos en los que se determinan una o más propiedades de uno o más miembros de una o más bibliotecas. La filtración puede realizarse computacionalmente utilizando modelos computacionales de biomoléculas y entorno virtual de las biomoléculas. En algunas realizaciones, se proporcionan sistemas de selección de proteínas virtuales para enzimas seleccionadas de actividad y selectividad deseadas.

[0091] Un "sistema de expresión" es un sistema para expresar una proteína o péptido codificado por un gen u otro ácido nucleico.

[0092] "Evolución dirigida", "evolución guiada," o "evolución artificial" se refiere a procesos *in silico, in vitro,* o *in vivo* para modificar artificialmente 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 reproductora en la cual (1) hay variedades de individuos, (2) algunas variedades tienen información genética hereditaria, y (3) algunas variedades difieren en su aptitud. El éxito reproductivo se determina por el resultado de la selección de una propiedad predeterminada tal 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.*

[0093] Los métodos de evolución dirigidos se pueden aplicar fácilmente a polinucleótidos para generar bibliotecas variantes que se pueden expresar, cribar y analizar. La mutagénesis y los métodos de evolución dirigida son bien conocidos en la técnica (véanse, por ejemplo, las patentes de los Estados Unidos números 5.605.793, 5.830.721, 6.132.970, 6.420.175, 6.277.638, 6.365.808, 6.602.986, 7.288.375, 6.287.861, 6.297.053, 6.576.467, 6.444.468, 5.811.238, 6.117.679, 6.165.793., 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, 6.426.224, 6.423.542, 6.479.652, 6.319.714, 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 et al., Meth. Mol. Biol., 57: 369 - 74 [1996]; Smith, Ann. Rev. Genet., 19: 423 - 462 [1985]; Botstein et al., Science, 229: 1193 - 1201 [1985]; Carter, Biochem. J., 237: 1-7 [1986]; Kramer et al., Cell, 38: 879 - 887 [1984]; Wells et al., Gene, 34: 315 - 323 [1985]; Minshull y otros, Curr. Op. Chem. Biol., 3: 284 - 290 [1999]; Christians et al., Nat. Biotechnol., 17: 259 - 264 [1999]; Crameri et al., Nature, 391: 288 - 291 [1998]; Crameri, et al., Nat. Biotechnol., 15: 436 - 438 [1997]; Zhang y col., Proc. Nat. Acad. Sci. EE.UU., 94: 4504 - 4509 [1997]; Crameri y otros, Nat. Biotechnol., 14: 315 - 319 [1996]; Stemmer, Nature, 370: 389 - 391 [1994]; Stemmer, Proc. Nat. Acad. Sci. EE.UU., 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).

[0094] En ciertas realizaciones, los métodos de evolución dirigida generan bibliotecas de variantes de proteínas por recombinación de genes que codifican variantes desarrolladas a partir de una proteína de matriz, así como por recombinación de genes que codifican variantes en una biblioteca variante de la proteína matriz. 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 variar por recombinación con otras variantes. El método se puede realizar durante uno o múltiples ciclos hasta que se logren los resultados deseados. Si se utilizan ciclos múltiples, cada uno típicamente implica una etapa 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 de selección de proteína virtual para determinar la actividad catalítica y la selectividad de enzimas para sustratos deseados.

[0095] En algunas realizaciones, los métodos de evolución dirigida generan variantes de proteínas por mutagénesis dirigida al sitio en los residuos definidos. Estos residuos definidos se identifican típicamente por 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 posibles (o tan cerca como posible) mutaciones en un sitio específico o región estrecha de un gen.

[0096] "Barajado" y "barajado de genes" son los 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 distintos de los fragmentos en sí mismos. Cada ciclo implica el recocido de fragmentos monocatenarios a través de la hibridación, el alargamiento posterior de los fragmentos recocidos a través de la extensión de la cadena y la desnaturalización. En el transcurso del barajado, una cadena creciente de ácido nucleico típicamente se expone a múltiples socios de apareamiento diferentes en un proceso denominado a veces "conmutación de molde", que implica cambiar un dominio de ácido nucleico de un ácido nucleico con un segundo dominio de un segundo nucleico ácido (es decir, ácidos nucleicos primero y segundo sirven como moldes en el procedimiento de barajado).

[0097] La conmutación de moldes 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 de molde durante los ciclos múltiples de recocido, extensión y desnaturalización. Por lo tanto, la transposición lleva 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 variantes múltiples). En algunas realizaciones de estas bibliotecas, las variantes contienen segmentos de secuencia de dos o más polinucleótidos parentales.

[0098] Cuando se emplean dos o más polinucleótidos parentales, los polinucleótidos parentales individuales son suficientemente homólogos que los fragmentos de diferentes padres se hibridan en las condiciones de recocido empleadas en los ciclos de barajado. En algunas realizaciones, el barajado 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 transposición puede producir polinucleótidos variantes muy diversos.

[0099] Diversas técnicas de barajado son conocidas en la técnica (véanse, por ejemplo, la Patente de los Estados Unidos N^{os} 6.917.882, 7.776.598, 8.029.988, 7.024.312, y 7.795.030).

[0100] Algunas de las técnicas de evolución dirigida emplean "empalme genético por extensión de solapamiento" o "empalme genético por extensión de solapamiento", 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 génicos superpuestos que se usan como ADN de molde 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 corte y empalme de genes. Las hebras 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 para fines de clonación (véanse, por ejemplo, Horton, et al., Biotechniques, 8 (5): 528-35 [1990]). "Mutagénesis" es el proceso de introducción de una mutación en una secuencia estándar o de referencia tal como un ácido nucleico principal o un polipéptido original.

[0101] La mutagénesis dirigida al sitio es un ejemplo de una técnica útil para introducir mutaciones, aunque cualquier método adecuado encuentra uso. De este modo, alternativamente o además, los mutantes pueden proporcionarse mediante síntesis génica, saturación de mutagénesis aleatoria, bibliotecas combinatorias semisintéticas de residuos, recombinación de secuencia recursiva ("RSR") (véanse, por ejemplo, publicación de solicitud de patente de Estados Unidos Nº 2006/0223143), barajado génico, PCR propensa a error, y/o cualquier otro método adecuado.

[0102] Un ejemplo de un procedimiento de mutagénesis de saturación adecuado se describe en la publicación de solicitud de patente de los Estados Unidos Nº 2010/0093560.

[0103] 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 polinucleótido. 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 el alargamiento de cadena parcial de ácido(s) nucleico(s) complementario(s). En algunas realizaciones que implican técnicas *in silico,* se generan fragmentos virtuales de forma computacional para imitar los resultados de fragmentos generados por técnicas químicas y/o biológicas. En algunas realizaciones, los fragmentos polipeptídicos exhiben la actividad del polipéptido de longitud completa, mientras que en algunas otras realizaciones, los fragmentos polipeptídicos no tienen la actividad exhibida por el polipéptido de longitud completa.

[0104] "Polipéptido parental", "polinucleótido parental", "ácido nucleico parental" y "progenitor" 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 tal como una evolución dirigida. En algunas realizaciones, el propio progenitor se produce a través de mezcla u otros procedimientos de generación de diversidad. En algunas realizaciones, los mutantes usados en la evolución dirigida están directamente relacionados con un polipéptido original. En algunas realizaciones, el polipéptido precursor es estable cuando se expone a condiciones de temperatura, pH y/o condiciones de disolvente y puede servir como base para generar variantes para la mezcla. En algunas realizaciones, el polipéptido parental no es estable en condiciones extremas de temperatura, pH y/o disolvente, y el polipéptido parental se desarrolla para producir variantes robustas.

[0105] Un "ácido nucleico original" codifica un polipéptido parental.

[0106] 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 ácidos nucleicos (por ejemplo, genes, oligonucleótidos, etc.) o productos de expresión (por ejemplo, enzimas u otras proteínas) a partir de los mismos. 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 1.000 moléculas diferentes. Para algunas aplicaciones, la biblioteca incluye al menos alrededor de 10.000 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 serie de ácidos nucleicos variantes o quiméricos o proteínas producidas por un procedimiento de evolución dirigida.

[0107] Dos ácidos nucleicos se recombinan cuando las secuencias de cada uno de los dos ácidos nucleicos se combinan para producir una progenie de ácido(s) nucleico(s). Dos secuencias se recombinan "directamente" cuando ambos ácidos nucleicos son sustratos para recombinación.

[0108] "Selección" se refiere al proceso en el que se identifican una o más biomoléculas que tienen 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 la filtración pueden ser, y a menudo son, simultáneas. Algunas realizaciones descritas en este documento proporcionan sistemas y métodos para la detección y selección de enzimas de actividad y/o selectividad deseables.

[0109] El término "modelo de secuencia-actividad" 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.

[0110] La "secuencia de referencia" es una secuencia a partir de la cual se efectúa la variación de la secuencia. En algunos casos, se usa una "secuencia de referencia" para definir las variaciones. Tal secuencia puede ser una 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 original de variantes de proteínas. En ciertas realizaciones, una secuencia de referencia es la secuencia de una proteína parental o ácido nucleico.

[0111] 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. Ejemplos de métodos adecuados de secuenciación de próxima generación incluyen, entre otros, secuenciación en una sola molécula en tiempo real (por ejemplo, Pacific Biosciences, Menlo Park, California), secuenciación de semiconductores iónicos (por ejemplo, Ion Torrent, South San Francisco, California), pirosecuenciación (*p. ej.,* 454, Branford, Connecticut), secuenciación por ligamiento (*p. ej.,* secuenciación SOLiD de Life Technologies, Carlsbad,

California), secuenciación por síntesis y terminador reversible (*p. ej.,* Illumina, San Diego, California), tecnologías de imágenes de ácido nucleico tales como microscopía electrónica de transmisión y similares.

[0112] Un "algoritmo genético" es un proceso que imita los procesos evolutivos. Los algoritmos genéticos (GA) se usan en una amplia variedad de campos para resolver problemas que no están completamente caracterizados o son demasiado complejos para permitir la caracterización completa, pero para los cuales se dispone de alguna evaluación analítica. Es decir, los GA se utilizan para resolver problemas que pueden evaluarse mediante una medida cuantificable del 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 descripción, 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 utilizados para entrenar un modelo, como un modelo de actividad de secuencia.

[0113] En una implementación típica, un algoritmo genético proporciona y evalúa una población de cadenas de caracteres en una primera generación. Una "función de aptitud física" evalúa a los miembros de la población y los clasifica en función de uno o más criterios, como la actividad alta. Las cadenas de caracteres de alto rango se seleccionan para promoción a una segunda generación y/o apareamiento para producir "cadenas de caracteres hijos" para la segunda generación. La población en la segunda generación es evaluada de manera similar por la función de aptitud, y los miembros de alto rango son promovidos y/o apareados como con la primera generación. El algoritmo genético continúa de esta manera para las generaciones posteriores hasta que se cumple un "criterio de convergencia", en cuyo punto el algoritmo concluye con uno o más individuos de alto rango.

[0114] El término "operación genética" (o "IR") se refiere a operaciones genéticas biológicas y/o computacionales, donde todos los cambios en cualquier población de cualquier tipo de cadenas de caracteres (y por lo tanto en cualquier propiedad física de objetos físicos codificados por tales cadenas) se puede describir como un resultado de la aplicación aleatoria y/o predeterminada de un conjunto finito de funciones algebraicas lógicas. Los ejemplos de GO incluyen, pero no están limitados a, multiplicación, cruce, recombinación, mutación, ligación, fragmentación, etc.

II. EXAMEN DE PROTEÍNAS VIRTUALES

[0115] En algunas realizaciones, un sistema de selección de proteína virtual está configurado para realizar diversas operaciones asociadas con la identificación computacional de variantes de biomoléculas que es probable que tengan una actividad deseable tal como catalizar de forma eficiente y selectiva una reacción a una temperatura definida. El sistema virtual de detección de proteínas puede tomar como entradas, representaciones de uno o más ligandos que están destinados a interactuar con las variantes. El sistema puede tomar como otras entradas, representaciones de las variantes de biomoléculas, o al menos los sitios activos 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. El sistema virtual de detección de proteínas puede aplicar información de atraque y restricciones de actividad para evaluar el funcionamiento de las variantes.

[0116] En ciertas realizaciones, un sistema de exploración de proteína virtual aplica una o más restricciones para distinguir posiciones activas e inactivas. Tales posiciones pueden ser generadas por un acoplador como se describió anteriormente o por otra herramienta. Se evalúa una posición de ligando en su entorno para determinar si una o más características del ligando están posicionadas en el entorno para dar como resultado una transformación catalítica u otra actividad definida. El entorno en cuestión es típicamente un sitio activo de una enzima u otra biomolécula.

[0117] Si se supone que un sustrato u otro ligando se une a un sitio activo de la biomolécula, la pregunta que debe hacerse es si se une de una manera "activa". Un programa de acoplamiento típico puede indicarle a uno si un ligando se unirá o no al sitio activo, pero no le dice a uno si se une de una manera "activa".

[0118] En ciertas realizaciones, la actividad se determina considerando una o más posiciones generadas por un acoplador u otra herramienta. Cada posición se evalúa para determinar si cumple con las restricciones asociadas con una actividad de interés (por ejemplo, una "actividad deseada"). Una posición activa es aquella en la que es probable que el ligando experimente una transformación catalítica o desempeñe algún papel deseado, tal como la unión covalente con el sitio de unión.

[0119] Cuando se considera el recambio catalítico de un sustrato como la actividad, el sistema de selección de proteína virtual puede configurarse para identificar posiciones que se sabe que están asociadas con una reacción particular. En algunas realizaciones, esto implica considerar un intermedio de reacción o un estado de transición en lugar del propio sustrato. Además del recambio, las posiciones 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, la conversión regioselectiva de productos, etc. En algunos casos, la actividad es unión covalente irreversible o reversible tal como inhibición covalente dirigida (TCI).

[0120] Las restricciones se pueden determinar de forma directa, manual, automática, empírica y/o en base a información previamente conocida. En un enfoque, un investigador evalúa el sitio activo y un sustrato nativo para

una proteína de tipo salvaje. Esto se debe a que se sabe que la proteína de tipo silvestre evolucionó para su sustrato nativo por naturaleza y, por lo tanto, tiene una constante catalítica óptima (k_{cat}). En algunos casos, las estructuras cristalinas de la proteína de tipo salvaje y el sustrato nativo o un complejo intermedio se han resuelto. La restricción se puede configurar en función del análisis estructural. Esto se conoce como un "enfoque directo" para determinar la restricción. En los casos en que tales estructuras cristalinas no están disponibles, la evaluación puede realizarse con un programa de acoplamiento, por ejemplo. Usando el programa, el investigador identifica restricciones asociadas con una transformación catalítica del sustrato nativo en la proteína de tipo salvaje. Esto se conoce como un enfoque manual o empírico para determinar las restricciones. En otro enfoque, las restricciones se determinan usando cálculos de mecánica cuántica. Por ejemplo, un investigador puede optimizar el sustrato o el estado intermedio o de transición en presencia de grupos funcionales de los residuos catalíticos (por ejemplo, Tyr) y/o cofactores (por ejemplo, NADHP), utilizando la mecánica cuántica y establecer la restricción para parecerse a esos estados. Este enfoque a veces se denomina enfoque automático o *ab initio.* Un ejemplo de una herramienta comercial que utiliza este enfoque es Gaussian disponible en www|.|Gaussian.com.

[0121] Las restricciones pueden tomar diversas formas. En ciertas realizaciones, algunas o todas estas restricciones son restricciones geométricas que especifican la(s) posición(es) relativa(s) de uno o más átomos en una posición de ligando en un espacio tridimensional. En algunas realizaciones, el espacio puede definirse con respecto a las posiciones de los átomos en un sitio activo.

[0122] Una "restricción geométrica" es una restricción que evalúa la geometría de dos o más restos participantes u otros elementos químicos. En ciertas realizaciones, uno de los participantes es un resto u otra especie química en el ligando. En algunas realizaciones, otro de los participantes es un resto u otra característica química de un sitio activo de una biomolécula. La fracción u otra característica química del sitio activo puede estar asociada con residuos en el sitio activo de la biomolécula (por ejemplo, una cadena lateral de residuo de aminoácido), una característica de un cofactor u otro compuesto que típicamente se asocia con el sitio activo y/o catálisis, y similares. Como ejemplo, en la reducción de cetonas por una proteína de cetoreductasa, el grupo carbonilo del sustrato puede ser un participante en una restricción geométrica y un resto de tirosina de un sitio activo de enzima puede ser un segundo participante en la restricción geométrica.

[0123] En general, las restricciones geométricas se hacen con respecto a un ligando por una parte y una o más características del entorno de unión por otra parte. En algunas realizaciones, el entorno puede incluir posiciones residuales de la cadena principal del péptido (o cadenas laterales) y/o cofactores u otros materiales no estructurales que normalmente residen en un sitio activo.

[0124] La geometría de los participantes en la restricción geométrica puede ser definida en términos de distancia entre restos, ángulos entre restos, relación de torsión entre restos, etc. A veces, una restricción incluye múltiples restricciones geométricas básicas usadas para caracterizar la actividad. Por ejemplo, una restricción sobre la posición de un sustrato puede definirse por las distancias entre dos o más pares de átomos. En la Figura 1 se muestra un ejemplo. En el caso de una relación de torsión, la restricción puede ser apropiada cuando un sustrato y una característica del entorno del sitio activo se ven como placas nominalmente paralelas que comparten un eje de rotación común. La posición angular relativa de estas placas alrededor del eje define la restricción de torsión.

[0125] La Figura 1 representa un ejemplo de un flujo de trabajo que puede emplearse para identificar restricciones geométricas para identificar posiciones activas. El flujo de trabajo representado asume que la enzima de tipo salvaje es una reductasa de cetona y el sustrato nativo es la acetofenona. Como se representa en la esquina superior izquierda de la Figura 1, la reacción nativa convierte la acetofenona en un alcohol correspondiente mediante catálisis estereoselectiva. La reacción introduce un centro quiral en el carbono acetilo del sustrato cetona. La reductasa de cetona de tipo salvaje controla la conversión de modo que solo se produce el enantiómero R. La reacción se lleva a cabo en presencia de NADPH como cofactor. La reacción se representa esquemáticamente en la esquina superior izquierda de la Figura 1.

[0126] En la esquina superior derecha de la Figura 1, se representa el mecanismo de catálisis y selectividad. Este mecanismo se considera al definir las restricciones geométricas utilizadas para distinguir las posiciones activas de las inactivas. Como parte del proceso, un investigador o sistema automatizado determina la orientación del sustrato de acetofenona con respecto a su entorno catalítico en la reductasa de cetona natural. En general, el entorno relevante incluye los residuos circundantes, los cofactores, etc. presentes cuando tiene lugar la transformación catalítica.

[0127] En el ejemplo representado, las características relevantes del entorno de sitio activo en la reductasa de cetona de tipo salvaje son las posiciones de los átomos en (1) un residuo de tirosina en la cadena principal de la enzima de tipo salvaje y (2) el cofactor NADPH. Otras características ambientales relevantes del sustrato en las posiciones activas son bolsillos secundarios dentro del sitio activo. Estos no se muestran en la Figura 1. Uno de los bolsillos secundarios acomoda el grupo fenilo del sustrato de acetofenona y otro acomoda el grupo metilo de la acetofenona. Juntos, estos bolsillos secundarios mantienen el sustrato en una orientación que dicta la estereoespecificidad de la reacción. En algunas realizaciones, la información anterior se recoge en base al análisis estructural de la estructura cristalina de la reductasa de cetona natural y del complejo de sustrato de acetofenona

natural. Por lo tanto, las restricciones geométricas se pueden definir directamente.

[0128] El mecanismo catalítico de cetorreductasa se representa por una secuencia de flechas mostradas en la disposición representada (esquina superior derecha de la Figura 1). Específicamente, el NADPH dona electrones a través de un ion hidruro que se acopla con el carbono carbonílico de la acetofenona. Al mismo tiempo, un par de electrones del oxígeno carbonílico de la acetofenona se dona al protón del residuo de tirosina, y un par de electrones del oxígeno hidroxílico de la tirosina se dona al protón del resto ribosa de NADP(H), completando así la conversión del sustrato al alcohol correspondiente. Como se observa, la reacción continúa mientras que el grupo fenilo del sustrato se mantiene en un sub-bolsillo más grande, su grupo metilo se mantiene en un sub-bolsillo secundario menor, y su grupo cetona se mantiene muy cerca hacia el grupo hidroxilo de tirosina.

[0129] Como se muestra adicionalmente en la Figura 1, la reductasa de cetona natural se desarrolla en una reductasa de cetona variante que cataliza estereoespecíficamente la conversión de un sustrato diferente, denominado "sustrato deseado" en este documento. Como se representa en el medio de la Figura 1, la reacción deseada es una conversión de cetona de terc-butilo de metilo al enantiómero S del alcohol correspondiente (1 alcohol etílico de terc-butilo). Se presume que la reacción está catalizada en un sitio activo de una enzima variante optimizada para la conversión y con el cofactor NADPH.

[0130] Para asegurar que la reacción se desarrolla con la estereoespecificidad deseada, se debe determinar una o más restricciones. Obsérvese que el sustrato nativo se convierte mediante la reductasa de cetona natural en el enantiómero R y el sustrato deseado se debe convertir mediante la variante en el enantiómero S. Por lo tanto, se puede considerar que el grupo tert-butilo del sustrato deseado debe colocarse en el bolsillo secundario que normalmente acomoda el grupo metilo del sustrato de acetofenona nativo y el grupo metilo del sustrato deseado debe colocarse en el bolsillo secundario que acomoda el grupo fenilo del sustrato nativo.

[0131] Con esto en mente, se puede definir un conjunto de restricciones posicionales como se representa en la esquina inferior izquierda de la Figura 1. Como se muestra allí, se definen diversas restricciones con respecto a la posición tridimensional del sustrato nativo tal como se encuentra en el sitio activo de la enzima WT en la estructura cristalina, con el fin de obtener el máximo volumen de negocios (k_{cat}). En otras palabras, la orientación del grupo funcional clave del sustrato nativo, incluido el carbono carbonilo y el oxígeno del carbonilo que determinan el recambio catalítico y cualquiera de los dos carbonos próximos al carbono del carbonilo que dicta la estereoselectividad, según se determina con respecto al diagrama en la esquina superior derecha de la Figura 1 se traduce en coordenadas X, Y, Z. Ya que los modelos de homología de todas las variantes se construyeron usando la estructura WT como molde, las coordenadas X, Y, Z son transferibles a las variantes. Con este marco de referencia, las posiciones del grupo funcional clave ($C_1(C_2)C=O$) del sustrato deseado se pueden comparar con las posiciones de los 4 átomos correspondientes del sustrato nativo, ya que se predice que se sientan en una orientación óptima hacia el residuo catalítico de tirosina y el cofactor NADPH. Es de destacar que los residuos para la unión de catálisis (p. ej., tirosina) y residuos para el cofactor (NADPH) se conservan en todas las variantes y solo se esperan cambios sutiles conformacionales o posicionales para esta tirosina y NADPH en todas las variantes. Con esto en mente, las restricciones posicionales representadas en la esquina inferior izquierda de la Figura 1 especifican un rango de posiciones del átomo de carbono de carbonilo del sustrato deseado, átomo de oxígeno de carbonilo y átomo de terc-butilo central con respecto a posiciones correspondientes del átomo de carbono de carbonilo del sustrato nativo, átomo de oxígeno de carbonilo y átomo de carbono de metilo. El rango de diferencias posicionales entre los átomos del sustrato deseado y los átomos correspondientes del sustrato nativo se representa mediante las distancias d1, d2 y d3. Como ejemplo, puede requerirse que cada una de estas distancias sea de 1 angstrom o más o menos para que una posición del sustrato deseado se considere una posición activa. Los valores de restricción generalmente se establecen para ser un rango que permite cierta flexibilidad que refleja los cambios conformacionales sutiles de la tirosina y el cofactor catalíticos en una variante. En algunas implementaciones, los criterios para estas distancias se refinan mediante algoritmos de aprendizaje automático.

[0132] En los ejemplos anteriores, las posiciones de los tres átomos relevantes del sustrato deseado se aproximan a las del sustrato nativo. Se espera que las variantes de cetoreductasa atracadas con el sustrato deseado en las posiciones que satisfagan las restricciones de posición anteriores sean catalíticamente activas y selectivas de S.

[0133] En general, el sistema de selección de proteína virtual puede aplicar restricciones geométricas de cualquiera de varios tipos. En algunas implementaciones, aplica la distancia absoluta entre los participantes. Por ejemplo, la distancia entre un átomo de oxígeno en el grupo de carbonilo de un sustrato y un átomo de un grupo de tirosina de un sitio activo puede especificarse como una restricción (por ejemplo, la distancia entre estos átomos debe ser 2 Å $\pm$ 0,5 Å). En otro ejemplo, el ángulo entre una línea definida por el eje entre los átomos de carbono y oxígeno en un grupo carbonilo y otra línea a lo largo de un eje de un grupo fenilo en un sitio activo es 120° $\pm$ 20°.

[0134] La parte inferior derecha de la Figura 1 representa ejemplos de tipos de restricciones geométricas, cada una definida entre uno o más átomos del sustrato deseado y uno o más átomos de la enzima o un cofactor (u otra entidad) dentro de un bolsillo de unión. Una restricción de *distancia* se define como la distancia entre un átomo en el sustrato y un átomo en un residuo de sitio activo, un cofactor, etc. En la restricción de *ángulo* se define para una posición por la relación angular entre dos o más ejes definidos en el sustrato y su entorno. Los ejes pueden ser

enlaces covalentes, líneas entre los átomos del sustrato y un resto en el bolsillo de unión, etc. Por ejemplo, se puede definir un ángulo entre un eje definido entre dos átomos en el sustrato y otro eje definido como la separación entre un átomo en un residuo y un átomo en el sustrato. En algunas otras realizaciones, un eje se define entre dos átomos en una cadena lateral de residuo y otro eje se define por la separación entre un átomo en el sustrato y un átomo en el residuo. En la esquina inferior derecha de la Figura 1 se representa un tipo adicional de restricción geométrica. Este tipo de restricción se denomina "restricción de torsión" y supone que dos entidades distintas en el bolsillo de encuadernación (una de las cuales suele ser todo o parte de el sustrato) comparten un eje de rotación común. La restricción de torsión puede definirse por un rango de posiciones angulares de una de las entidades con respecto a la otra alrededor del eje de rotación común.

[0135] En general, la restricción geométrica puede aplicarse con respecto a alguna posición u orientación geométrica preestablecida de un resto de sustrato dentro de un bolsillo de unión. Dicha posición u orientación puede especificarse mediante, por ejemplo, una posición representativa de un resto activo en un sustrato nativo en un bolsillo de unión. Como ejemplo, los átomos de carbono y oxígeno del grupo de carbonilo del sustrato considerado deben estar dentro de 1 Å de las ubicaciones de los átomos de oxígeno de carbono de un grupo carbonilo en un sustrato nativo en el bolsillo de unión. Vea la restricción posicional que se muestra en la esquina inferior izquierda de la Figura 1. Tenga en cuenta que las restricciones posicionales en la esquina inferior izquierda de la Figura 1 existen entre el sustrato deseado y el sustrato nativo. Sin embargo, las restricciones posicionales se pueden traducir en relaciones entre el sustrato deseado y las variantes de la enzima, que corresponden a las restricciones geométricas en la esquina inferior central y derecha de la Figura 1.

[0136] Además de determinar las restricciones geométricas de forma directa, manual o automática utilizando sistemas informáticos, las restricciones también pueden refinarse mediante los resultados de la detección. Por ejemplo, si una o más de una variantes se identifican como activas mientras que otras se identifican como inactivas para la reacción deseada a través del examen de laboratorio, sus posiciones se pueden analizar más a fondo y se pueden entrenar las restricciones.

[0137] Mientras que el ejemplo representado en la Figura 1 utiliza una molécula relativamente pequeña y simple (cetona de terc-butilo de metilo) como un sustrato deseado, sustratos mucho más grandes y más complejos a menudo se evaluaron en un esfuerzo de evolución dirigida.

[0138] La Figura 2 presenta un flujo de trabajo para analizar la actividad potencial de las biomoléculas candidatas en algunas implementaciones. Si bien se pueden considerar muchas actividades diferentes, la que se enfatizará en esta realización es la transformación catalítica del sustrato. La transformación puede ser enantioselectiva o regioselectiva. En tal caso, las variantes son enzimas. En la descripción de esta figura, cuando se usa el término "sustrato", el concepto se extiende a ligandos relacionados tales como intermedios de reacción o estados de transición que son importantes en un paso de determinación de velocidad en la transformación catalítica del sustrato a un producto de reacción.

[0139] Como se muestra en la Figura 2, el proceso comienza identificando restricciones para distinguir posiciones activas de las inactivas del sustrato. Véanse bloque 201. En algunos casos, las restricciones se identifican por acoplamiento. En dichos procesos, un investigador toma en consideración la interacción del sustrato o el intermedio de reacción o estado de transición con el sitio activo de la enzima. En el proceso, identifica restricciones que dan como resultado la actividad deseada (por ejemplo, transformación catalítica estereoespecífica del sustrato). El investigador puede hacer esto con la ayuda del análisis de estructura, un programa de acoplamiento y/o cálculos de mecánica cuántica que presentan una representación de una enzima y un sustrato asociado, intermedio o estado de transición. El acoplamiento realizado con un acoplador a veces se conoce como un enfoque de acoplamiento "empírico" y la optimización realizada con una herramienta de mecánica cuántica a veces se denomina enfoque "*ab initio*". En algunas realizaciones, el acoplamiento se realiza con una enzima de tipo salvaje y el sustrato nativo, intermedio o estado de transición. Véanse el bloque 201. Como se explicó anteriormente, algunas restricciones son restricciones geométricas que representan las posiciones relativas de restos en los sustratos y restos deseados en el sustrato nativo o un cofactor asociado como se muestra en la esquina inferior izquierda de la Figura 1. En algunas implementaciones, las restricciones se pueden definir como las relaciones entre los sustratos deseados y las variantes de la enzima, como las restricciones geométricas que se muestran en la esquina inferior central y derecha de la Figura 1.

[0140] En algunos casos, las restricciones para posiciones activas se pueden identificar mediante técnicas distintas de atracar un sustrato nativo en una enzima de tipo salvaje. Por ejemplo, es posible identificar restos relevantes para una reacción catalítica y definir relaciones entre los restos identificados usando mecanismos de mecánica cuántica y dinámica molecular.

[0141] Volviendo al proceso que se muestra en la Figura 2, el sistema de selección de proteína virtual crea o recibe modelos estructurales para cada una de las múltiples biomoléculas variantes que deben considerarse para la actividad. Véanse el bloque 203. Como se explicó, los modelos estructurales son representaciones tridimensionales producidas computacionalmente de los sitios activos u otros aspectos de las variantes de la enzima. Estos modelos pueden guardarse para su uso posterior en una base de datos u otro repositorio de datos. En algunos casos, al

menos uno de los modelos se crea para su uso en el flujo de trabajo. En algunos casos, al menos uno de los modelos se creó previamente, en cuyo caso el proceso simplemente recibe dichos modelos.

[0142] Múltiples modelos, cada uno para una secuencia de biomolécula diferente se utilizan en el proceso mostrado en la Figura 2. Esto debe ser contrastado con flujos de trabajo convencionales utilizando programas de atraque. Los flujos de trabajo convencionales se centran en un único objetivo o secuencia. En algunos casos, un flujo de trabajo convencional considera múltiples instancias de un receptor, pero éstas se basan en la misma secuencia. Cada una de las instancias tiene diferentes coordenadas tridimensionales generadas a partir de simulaciones de RMN o dinámica molecular.

[0143] Los modelos estructurales utilizados en el proceso de la Figura 2 pueden variar de unos a otros por la inserción, deleción o sustitución en los modelos de uno o más residuos de aminoácidos en las posiciones asociadas con el sitio activo o con alguna otra posición en la enzima de secuencia. Los modelos estructurales pueden ser creados por varias técnicas. En una realización, se crean por modelado de homología.

[0144] Con las limitaciones de actividad y los modelos estructurales en su lugar, el sistema virtual de detección de proteínas itera las variantes que han sido seleccionadas para su consideración. El control de la iteración se ilustra mediante un bloque 205, que indica que la siguiente enzima variante bajo consideración se selecciona para el análisis. Esta operación y las operaciones restantes de la Figura 2 pueden implementarse mediante software o lógica digital.

[0145] Para la enzima variante actualmente en consideración, el sistema de selección de proteína virtual primero intenta acoplar el sustrato deseado al sitio activo de la variante. Véanse bloque 207. Este proceso puede corresponder a un procedimiento convencional de acoplamiento. Por lo tanto, se puede emplear un acoplador para determinar si el sustrato es capaz o no de atracar con el sitio activo en la variante. Esta decisión se representa en un bloque 209. Tenga en cuenta que el sustrato deseado a veces es diferente del sustrato nativo, que puede haberse utilizado para generar las restricciones.

[0146] Si el sistema de exploración de proteína virtual determina que es poco probable que el acoplamiento sea exitoso, el control del proceso se dirige a un bloque 220, donde el sistema determina si hay otras variantes a considerar. Si no hay otras variantes a tener en consideración, el proceso se completa con una operación opcional 223, como se indica. Si, por otro lado, una o más variantes quedan por considerar, el control del proceso se dirige de vuelta al paso de proceso 205 donde se selecciona la siguiente variante para consideración. Esta variante se evalúa luego por su capacidad de atracar el sustrato considerado como se describió anteriormente con referencia a los bloques 207 y 209.

[0147] Si resulta que la variante bajo consideración puede acoplarse con éxito al sustrato, el control del proceso se dirige a una parte del algoritmo donde se consideran múltiples posiciones y cada una se evalúa para determinar su actividad. Como se describe a continuación, este análisis se representa por los bloques 211, 213, 215, y 217.

[0148] Como se muestra, se repite el proceso a través de múltiples posiciones disponibles. En diversas realizaciones, un acoplador ayuda a seleccionar las posiciones. Como se explicó, los estibadores pueden generar numerosas posiciones de un sustrato en un sitio activo. También puede clasificar las posiciones según uno o más criterios, como la puntuación de atraque, las consideraciones energéticas, etc. Se puede considerar la energía total y/o la energía de interacción, como se describe en otra parte. Independientemente de cómo se generan y/o clasifican las posiciones, el flujo de trabajo puede configurarse para considerar un número específico de posiciones. El número de posiciones que se considerarán se puede establecer arbitrariamente. En una realización, se consideran al menos aproximadamente las 10 posiciones más importantes. En otra realización, se consideran al menos aproximadamente 20 posiciones, o al menos aproximadamente 50 posiciones, o al menos aproximadamente 100 posiciones. Sin embargo, no se pretende que la presente invención se limite a un número específico de posiciones.

[0149] Como se representa en el bloque 211, el proceso selecciona la siguiente posición para el análisis. La posición seleccionada actualmente se evalúa luego contra las restricciones identificadas en el bloque 201, para determinar si la posición es una posición activa. Como se explicó, tales restricciones pueden ser restricciones geométricas que determinan si uno o más restos del sustrato están ubicados dentro del sitio activo, de manera que es probable que el sustrato experimente una transformación catalítica deseada.

[0150] Si la evaluación realizada en el bloque 213 indica que la posición actual no es una posición activa, el sistema de exploración de proteína virtual determina entonces si hay otras posiciones a considerar para la variante actual bajo consideración. Véanse el bloque 215. Suponiendo que hay más posiciones que considerar, el control del proceso se dirige de nuevo al bloque 211, donde se considera la siguiente postura.

[0151] Suponiendo que el sistema virtual de selección de proteína determina en el bloque 213 que la posición considerada está activa, toma nota de esta posición para su posterior consideración. Véanse el bloque 217. En algunas realizaciones, el sistema de selección de proteína virtual puede mantener una cuenta corriente del número

de posiciones activas para la variante actualmente bajo consideración.

[0152] Después de señalar apropiadamente que la posición actual está activa, el control del proceso se dirige al bloque 215, donde el sistema de exploración de proteína virtual determina si hay otras posiciones a considerar. Después de repetir la consideración de todas las posiciones disponibles para la variante bajo consideración, el sistema virtual de detección de proteínas determina que no hay más posiciones que considerar y el control del proceso se dirige a un bloque 218, que caracteriza la actividad probable de la variante actual. La caracterización se puede realizar mediante diversas técnicas, que incluyen, pero no se limitan al número de posiciones activas y puntajes de acoplamiento asociados para la variante en consideración y otras consideraciones como se describen en este documento. Después de que se completa la operación del bloque 218, el control del proceso se dirige a la operación de decisión 220, que determina si hay otras variantes a considerar. Si hay variantes adicionales a considerar, el control del proceso se devuelve al bloque 205, donde el flujo de trabajo continúa como se describió anteriormente.

[0153] Después de considerar todas las variantes en el flujo de trabajo, el sistema virtual de detección de proteínas puede clasificarlas según uno o más criterios, como el número de posiciones activas que tienen las variantes, una o más puntuaciones de atraque de las posiciones activas, y/o una o más energías de enlace de las posiciones activas. Véanse bloque 223. Solo las posiciones identificadas como posiciones activas (bloque 217) necesitan ser evaluadas al realizar el ranking del bloque 223. De esta forma, las operaciones en el flujo de trabajo sirven para filtrar posiciones inactivas de posiciones activas y ahorrar esfuerzo computacional asociado con la clasificación de las variantes. Si bien no se muestra en la Figura 2, las variantes se pueden seleccionar para una mayor investigación en función de su clasificación.

[0154] En ciertas realizaciones, se ejecuta un protocolo para calcular energías de enlace para evaluar las energías de cada posición 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 estibadores. Se encuentran disponibles varios modelos de solvatación para calcular energías de enlace, que incluyen, pero no se limitan a dieléctricos dependientes de la distancia, generalizado nacido con suma por pares (GenBorn), generalizado nacido with membrana implícita (GBIM), generalizado nacido con integración de volumen molecular (GBMV), generalizado nacido con una conmutación 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 separados de programas de acoplador. Generalmente producen resultados que son más precisos que los puntajes de atraque, debido en parte a la inclusión de efectos de solvatación en sus cálculos. En varias implementaciones, las energías de enlace se calculan solo para las posiciones que se consideran activas.

A. Generación de modelos de múltiples biomoléculas que contienen cada una un sitio activo

[0155] Un sistema informático puede proporcionar modelos tridimensionales para una pluralidad de variantes de proteínas. Los modelos tridimensionales son representaciones computacionales de algunas o todas las secuencias completas de las variantes de proteínas. Típicamente, como mínimo, las representaciones de cálculo cubren al menos los sitios activos de las variantes de proteínas.

[0156] En algunos casos, los modelos tridimensionales son modelos de homología preparados usando un sistema informático diseñado apropiadamente. Los modelos tridimensionales emplean un molde estructural en el que las variantes de proteínas varían entre sí en sus secuencias de aminoácidos. En general, un molde estructural es una estructura previamente resuelta mediante 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 del molde 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.

[0157] Los modelos tridimensionales de las variantes de proteínas se pueden producir mediante técnicas distintas de la modificación de homología. Un ejemplo es el enhebrado de proteínas, que también requiere un molde de estructura. Otro ejemplo es el *modelado de* proteínas *ab initio* o *de novo,* que no requiere un molde 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.

[0158] En algunas realizaciones, las variantes de proteína varían entre sí en sus sitios activos. En algunos casos, los sitios activos se diferencian entre sí por al menos una mutación en la secuencia de aminoácidos del sitio activo. La(s) mutación(es) pueden realizarse en una secuencia de proteína de tipo salvaje o en 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 activo pero difieren en la secuencia de aminoácidos para otra región de la proteína. En algunos casos, dos variantes de proteína se diferencian entre sí por 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 un número específico de diferencias de aminoácidos entre variantes de proteínas.

[0159] 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 usadas en la evolución dirigida incluyen barajado de genes, mutagénesis, recombinación y similares. Ejemplos de técnicas de evolución dirigidas se describen en la publicación de solicitud de patente de los Estados Unidos Nº 2006/0223143.

[0160] En algunos procesos, la pluralidad de variantes incluye al menos aproximadamente diez variantes diferentes, o al menos aproximadamente 100 variantes diferentes, o al menos aproximadamente mil variantes diferentes.

B. Evaluación de un ligando en múltiples variantes de proteínas diferentes

[0161] Como se explica en la presente memoria, el acoplamiento se lleva a cabo mediante un sistema informático apropiadamente programado que utiliza una representación computacional de un ligando y representaciones computacionales de los sitios activos de la pluralidad generada de variantes.

[0162] Como ejemplo, un acoplador se puede configurar para realizar algunas o todas las siguientes operaciones:

1. Generar un conjunto de conformaciones de ligandos 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 que se generarán se puede establecer de forma arbitraria. 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. Generar 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 puede establecerse 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 un 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 suposiciones físicamente poco realistas sobre la permisibilidad de ciertas orientaciones en un sitio activo. Por ejemplo, el acoplador puede suponer que los átomos del ligando y los átomos del sitio activo pueden ocupar esencialmente el mismo espacio, lo cual es imposible en función de la repulsión de Pauli y las consideraciones estéricas. Esta suposición suavizada puede implementarse, por ejemplo, empleando una forma relajada del potencial de Lennard-Jones al explorar el espacio de conformación. Al usar un cálculo de energía suavizado, el acoplador permite una exploración más completa de las conformaciones que las disponibles utilizando consideraciones de energía físicamente realistas. Si la energía ablandada 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 conservan como "posiciones". En ciertas implementaciones, este proceso continúa hasta que se encuentra un número deseado de posiciones de baja energía o se encuentra un número máximo de posiciones malas.
3. Someter cada posición retenida del paso 2 a dinámicas moleculares recocidas simuladas para refinar la posición. La temperatura 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 físicamente más realista que la proporcionada por el cálculo de energía suavizada.
4. Realizar una minimización final del ligando en el receptor rígido utilizando potencial no ablandado. Esto proporciona un valor de energía más preciso para las posiciones retenidas. Sin embargo, el cálculo puede proporcionar solo información parcial sobre las energías de las posiciones.
5. Para cada posición final, calcular 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 posiciones se ordenan por energía CHARMm y se retienen las posiciones de puntuación superior (la mayoría de las negativas, por lo tanto, favorables para el encuadernado). En algunas realizaciones, este paso (y/o paso 4) elimina posiciones que son energéticamente desfavorables.

[0163] La siguiente referencia proporciona un ejemplo de funcionamiento de un descargador: 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).

[0164] Un acoplador tal como el descrito aquí puede proporcionar una o más piezas de información usadas por el sistema de exploración para identificar variantes de alto rendimiento. Dicha información incluye la identidad de variantes para las que es improbable el acoplamiento con el sustrato deseado. Esas variantes no necesitan evaluación para la actividad, etc. Otra información proporcionada por el acoplador incluye conjuntos de posiciones (un conjunto para cada variante) que se pueden considerar para la actividad. Otra información más incluye puntajes de atraque de las posiciones en los conjuntos.

C. Determinar si las posiciones del ligando acoplado son activas

[0165] Para una variante de proteína que se une satisfactoriamente con el ligando, el sistema de selección de proteína virtual realiza las siguientes operaciones: (i) considerar una pluralidad de posiciones de la representación computacional del ligando en el sitio activo de la variante de proteína en consideración, y (ii) determinar cuál de las varias posiciones está activa.

[0166] Una posición activa es una que reúne una restricción más para que el ligando se una bajo condiciones definidas (en lugar de una condición de unión arbitraria). Si el ligando es un sustrato y la proteína es una enzima, la unión activa puede ser un enlace que permita que el sustrato experimente una transformación química catalizada, particularmente una transformación estereoespecífica. En algunas implementaciones, las restricciones son restricciones geométricas que definen un rango de posiciones relativas de uno o más átomos en el ligando y uno o más átomos en la proteína y/o cofactor asociados con la proteína.

[0167] En algunos casos, las restricciones se identifican a partir de una o más conformaciones de un sustrato nativo y/o un intermedio posterior cuando se somete a una transformación química catalizada por una enzima de tipo salvaje. En ciertas realizaciones, las restricciones incluyen (i) una distancia entre un resto particular en el sustrato y/o el intermedio posterior y un resto particular o resto en el sitio activo, (ii) una distancia entre un resto particular en el sustrato y/o un intermedio posterior y un cofactor particular en el sitio activo, y/o (iii) una distancia entre un resto particular en el sustrato y/o el intermedio posterior y un resto particular en un sustrato nativo posicionado idealmente, y/o intermedio subsiguiente en el sitio activo. En ciertas realizaciones, las restricciones pueden incluir ángulos entre enlaces químicos, torsión alrededor de ejes o tensión en enlaces químicos.

[0168] La pluralidad de posiciones de la representación computacional del sustrato y/o el intermedio posterior se puede generar con respecto a una representación computacional de la variante de proteína en consideración. La pluralidad de posiciones se puede generar mediante diversas técnicas. Los ejemplos generales de tales técnicas incluyen búsquedas torsionales sistemáticas o estocásticas sobre enlaces giratorios, simulaciones de dinámica molecular y algoritmos genéticos diseñados para localizar conformaciones de baja energía. En un ejemplo, las posiciones se generan utilizando una dinámica molecular de alta temperatura, seguida de rotación aleatoria, refinamiento mediante recocido simulado basado en rejilla y una minimización final de campo de fuerza o de rejilla para generar una conformación y/u orientación del sustrato y/o intermedio posterior en el sitio activo de representación computacional. Algunas de estas operaciones son opcionales, por ejemplo, el refinamiento mediante el recocido simulado basado en la red y la minimización del campo de fuerza o de la cuadrícula.

[0169] En ciertas realizaciones, el número de posiciones consideradas 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 posiciones consideradas.

[0170] Si el proyecto tiene éxito, se determina que al menos una de las variantes tiene una o más posiciones que son activas y energéticamente favorables. En ciertas realizaciones, una variante seleccionada para consideración adicional es una determinada que tiene un gran número de conformaciones activas en comparación con otras variantes. En ciertas realizaciones, las variantes son seleccionadas clasificando las variantes según el número de posiciones activas que tienen, una o más puntuaciones de atraque para las posiciones activas, y/o una o más energías vinculantes de las posiciones activas. Como ejemplos, los tipos de puntajes de atraque que se pueden considerar incluyen los puntajes basados en la fuerza de van de Waals y/o interacción electrostática. Como ejemplos, los tipos de energías de enlace que se pueden considerar incluyen la fuerza de van der Waals, la interacción electrostática y la energía de solvatación.

[0171] Una variante de proteína determinada para soportar una o más posiciones activas puede seleccionarse para investigación adicional, síntesis, producción, etc. En un ejemplo, se usa una variante de proteína seleccionada para sembrar una o más rondas de evolución dirigida. Como ejemplo, una ronda de evolución dirigida 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 usando la pluralidad de oligonucleótidos. Los oligonucleótidos se pueden preparar por cualquier medio adecuado, que incluye, pero no se limita a, síntesis génica, fragmentación de un ácido nucleico que codifica una parte o la totalidad de la variante de proteína seleccionada, etc. En ciertas realizaciones, la ronda de evolución dirigida incluye fragmentar y recombinar la pluralidad de oligonucleótidos. En ciertas realizaciones, la ronda de evolución dirigida incluye realizar mutagénesis de saturación en la pluralidad de oligonucleótidos.

[0172] Las transformaciones químicas catalizadas que pueden cribarse usando restricciones incluyen, pero no están limitadas a, por ejemplo, reducción de cetona, transaminación, oxidación, hidrólisis de nitrilo, reducción de imina, reducción de enona, hidrólisis de acilo y deshalogenación de halohidrina. Los ejemplos de clases de enzimas que pueden proporcionar las múltiples variantes evaluadas utilizando restricciones incluyen, pero no se limitan a: reductasa de cetonas, transaminasas, citocromo P450, moooxigenasas Baeyer-Villiger, monoaminooxidasas, nitrilasa, reductasas de imina, reductasas de enona, acilasas y deshidrogenasas de halohidrina. En el contexto del

diseño racional de ligandos, la optimización de la inhibición covalente dirigida (TCI) es un tipo de actividad que se puede cribar para usar restricciones. Un ejemplo de una aplicación de TCI se describe en Singh et al., The resurgence of covalent drugs, Nature Reviews Drug Discovery, vol. 10, pp. 307-317 (2011). En algunas implementaciones, la actividad de TCI se encuentra identificando un aminoácido nucleófilo (por ejemplo, cisteína) en una proteína. El proceso descrito en este documento puede ayudar a identificar inhibidores que satisfagan restricciones que definen una orientación ideal de un resto electrófilo importante para la inhibición (un inhibidor putativo) que puede reaccionar con la biomolécula que se va a inhibir.

III. UTILIZAR EL SISTEMA DE SELECCIÓN DE PROTEÍNA VIRTUAL PARA DISEÑAR ENZIMAS

[0173] Algunas realizaciones proporcionan procesos para virtualmente modelar y seleccionar enzimas usando un sistema de selección de proteína virtual, identificando de ese modo enzimas que tienen propiedades deseadas, por ejemplo, actividad catalítica y selectividad. En algunas realizaciones, una familia de enzimas reales se puede modelar virtualmente y cribar como una biblioteca de variantes inicial. Algunas realizaciones pueden usar iterativamente una o más enzimas seleccionadas mediante cribado virtual de la biblioteca inicial como polipéptidos parentales o secuencias de referencia para generar una nueva biblioteca variante mediante técnicas *in silico, in vitro* o *in vivo.* En algunas realizaciones, una o más enzimas altamente clasificadas por el sistema como se describe en este documento se seleccionan como polipéptido(s) original(es). La nueva biblioteca variante incluye secuencias de proteínas que son diferentes de las secuencias de los polipéptidos originales, y/o pueden usarse como precursores para introducir variaciones posteriores.

[0174] En algunas realizaciones, los polipéptidos parentales se modifican en un procedimiento de evolución dirigida realizando mutagénesis y/o un mecanismo de generación de diversidad basado en recombinación para generar la nueva biblioteca de variantes de proteína. En algunas realizaciones, los polipéptidos precursores se alteran mediante al menos una sustitución, inserción, cruzamiento, deleción y/u otra operación genética. La evolución dirigida puede implementarse directamente sobre los polipéptidos (por ejemplo, en un proceso *in silico)* o indirectamente sobre los ácidos nucleicos que codifican los polipéptidos (por ejemplo, en un proceso *in vitro)*. La nueva biblioteca se puede usar para generar nuevos modelos de homología para un análisis posterior y una evolución dirigida.

[0175] En algunas realizaciones, el modelado, rastreo y evolución de las enzimas se llevan a cabo iterativamente *en silico* hasta que se cumplen una o más enzimas que cumplen ciertos criterios. Por ejemplo, los criterios pueden ser una energía o puntaje de unión especificado, o una mejora de los mismos. Otras realizaciones pueden combinar técnicas *in silico* y físicas (p. ej., *In vitro* o *in vivo)*. Por ejemplo, es posible iniciar un proceso de diseño de enzimas utilizando enzimas derivadas de cribado *in vitro* y secuenciación. *La* secuenciación *in vitro* se puede realizar mediante secuenciación de próxima generación. Luego, el proceso de diseño de la enzima puede usar métodos *in silico* para la evolución dirigida, el modelado y la detección posterior. El proceso finalmente puede usar técnicas *in vitro* y/o *in vivo* para validar una enzima en un sistema biológico. Otras combinaciones y órdenes de técnicas *in silico* y físicas son adecuadas para diversas aplicaciones. De hecho, no se pretende que la presente invención se limite a ninguna combinación específica y/u orden de métodos.

[0176] En algunas realizaciones, la preparación de secuencias polipeptídicas se logra *in silico.* En otras realizaciones, los polipéptidos se generan sintetizando oligonucleótidos o secuencias de ácido nucleico usando un sintetizador de ácido nucleico y traduciendo las secuencias de nucleótidos para obtener los polipéptidos.

[0177] Como se indicó anteriormente, en algunas realizaciones, la enzima seleccionada puede modificarse realizando uno o más mecanismos de generación de diversidad basados en la recombinación para generar la nueva biblioteca de variantes de proteínas. Tales mecanismos de recombinación incluyen, pero no están limitados a, por ejemplo, barajado, cambio de molde, empalme de genes mediante extensión de solapamiento, PCR propensa a errores, bibliotecas combinatorias semisintéticas de residuos, recombinación de secuencia recursiva ("RSR") (véanse, por ejemplo, Publicación de Solicitud de Patente de Estados Unidos Nº 2006/0223143). En algunas realizaciones, algunos de estos mecanismos de recombinación pueden implementarse *in vitro.* En algunas realizaciones, algunos de estos mecanismos de recombinación pueden implementarse computacionalmente *en silico* para imitar los mecanismos biológicos.

[0178] Algunas realizaciones incluyen la selección de una o más posiciones en una secuencia de proteína y llevar a cabo métodos de mutación dirigidos al sitio tales como mutagénesis de saturación en una o más posiciones así seleccionadas. En algunas realizaciones, las posiciones se seleccionan evaluando la estructura del sitio activo y/o restricciones relacionadas con la reacción catalítica como se discute en otra parte del documento. La combinación de selección virtual con modelado de actividad de secuencia encuentra uso en algunas realizaciones. En estas realizaciones, el proceso de evolución dirigida puede seleccionar las posiciones evaluando los coeficientes de los términos de un modelo de actividad de secuencia, identificando de ese modo uno o más de los residuos que contribuyen a la actividad de interés. La Patente de Estados Unidos Nº 7.783.428 proporciona ejemplos de modelos de actividad de secuencia que pueden usarse para identificar aminoácidos para la mutagénesis.

[0179] En algunas realizaciones, el método implica seleccionar uno o más miembros de la nueva biblioteca de

variantes de proteínas para producción. Una o más de estas variantes pueden entonces sintetizarse y/o expresarse en un sistema de expresión. En una realización específica, el método continúa de la siguiente manera: (i) proporcionar un sistema de expresión a partir del cual se puede expresar un miembro seleccionado de la nueva biblioteca de variantes de proteínas; y (ii) expresar el miembro seleccionado de la nueva biblioteca de variantes de proteínas.

[0180] Las Figuras 3A-3C son diagramas de flujo que muestran ejemplos de flujos de trabajo para diseñar secuencias de biomoléculas, que implementan diversas combinaciones de elementos descritos en este documento. La Figura 3A muestra un diagrama de flujo para un proceso 300 que comienza recibiendo información de secuencia de múltiples secuencias de partida de un panel de biomoléculas, tal como un panel de enzimas. Véanse el bloque 302. El proceso luego realiza un cribado virtual de las secuencias recibidas actualmente utilizando un sistema virtual de detección de proteínas. Véanse el bloque 304. En algunas realizaciones, el sistema de selección de proteína virtual puede crear modelos de homología tridimensional de las secuencias de inicio y acoplar uno o más sustratos con los modelos de homología considerando las posiciones de los sustratos como se describió anteriormente, generando así puntajes de atraque para las secuencias de inicio. El sistema virtual de detección de proteínas también puede calcular la energía de interacción y la energía interna de los participantes en el atraque (las enzimas y los sustratos). Además, el sistema de selección de proteína virtual puede evaluar diversas restricciones de posiciones para determinar si las posiciones son activas, es decir, los sustratos se unen con la enzima de una manera que es probable que provoque una conversión catalítica del sustrato. Además, en algunas realizaciones, la evaluación de las restricciones también proporciona inferencia con respecto a si los productos de la reacción catalítica son enantioselectivos y/o regioselectivos. En algunas realizaciones, el proceso selecciona una o más secuencias basadas en la energía de unión, la actividad y la selectividad determinadas por el sistema de exploración virtual. Véanse el bloque 306. El proceso luego evalúa si es necesario llevar a cabo una investigación adicional de las secuencias seleccionadas en el paso 308. Si es así, el proceso en este ejemplo muta computacionalmente las secuencias seleccionadas. Las mutaciones se basan en los diversos mecanismos de generación de diversidad descritos anteriormente, como la mutagénesis o la recombinación. Véanse el bloque 310. A continuación, se proporcionan las secuencias mutadas computacionalmente para una nueva ronda de exploración virtual mediante el sistema virtual de detección de proteínas. Véanse el bloque 304. La selección y selección virtual puede continuar para las iteraciones, hasta que no sea necesaria una investigación adicional de las secuencias, que puede determinarse mediante criterios preestablecidos, tales como un número específico de iteraciones y/o un nivel particular de actividad deseada. En ese punto, el proceso de diseño de biomoléculas (por ejemplo, enzimas) se termina en el paso 312.

[0181] La Figura 3B muestra un diagrama de flujo para un proceso 320 para la evolución dirigida de biomoléculas tales como enzimas, cuyo proceso tiene algunos elementos similares y algunos diferentes en comparación con el proceso de 300. El proceso 320 comienza por síntesis *in vitro* de múltiples secuencias de inicio de biomoléculas (por ejemplo, enzimas), que pueden ser necesarias o útiles cuando un panel preexistente de biomoléculas no está disponible. Véanse el bloque 322. Las secuencias sintetizadas también se pueden ensayar para recopilar datos para las secuencias, datos que pueden ser útiles para diseñar biomoléculas de propiedades deseadas, en las que los datos no pueden obtenerse mediante el sistema de exploración virtual. El proceso luego realiza un cribado virtual de las secuencias sintetizadas usando un sistema de cribado de proteínas virtual, representado en el bloque 324, que es similar al paso 304 en el proceso 300. El proceso selecciona una o más secuencias basadas en la energía de enlace, la actividad y selectividad determinada por el sistema de evaluación virtual. Véanse el bloque 326. El proceso luego evalúa si es necesario realizar una evolución dirigida adicional de las secuencias seleccionadas en el paso 328. Si es así, el proceso en este ejemplo muta las secuencias seleccionadas *in silico* o *in vitro.* Las mutaciones se basan en los diversos mecanismos de generación de diversidad descritos anteriormente. Véanse el bloque 330. Las secuencias mutadas se proporcionan a continuación para una nueva ronda de cribado virtual mediante el sistema virtual de cribado de proteínas. Véanse el bloque 324. El cribado y la selección virtuales pueden continuar para las iteraciones, hasta que no sean necesarias más evoluciones de secuencias, que pueden determinarse mediante criterios preestablecidos, tales como un número específico de iteraciones y/o un nivel particular de actividad deseada. En ese punto, las secuencias seleccionadas por el sistema de selección virtual se sintetizan y expresan para producir enzimas reales. Véanse el bloque 332. Las enzimas producidas se pueden analizar para actividades de interés, que se pueden usar para validar los resultados del proceso de selección virtual. Véanse el bloque 334. Después del ensayo, el proceso de evolución dirigido se concluye en el paso 336.

[0182] La Figura 3C muestra un diagrama de flujo para un proceso 340 para la evolución dirigida de biomoléculas tales como enzimas. El proceso 340 comienza por una evolución dirigida *in vitro* para derivar múltiples secuencias de partida de biomoléculas (por ejemplo, enzimas). Véanse el bloque 342. Como en el proceso 320, las secuencias derivadas se analizan para determinar si las secuencias cumplen ciertos criterios, tales como actividad o selectividad deseadas. Las secuencias que cumplen los criterios se determinan como aciertos para un mayor desarrollo. Véanse el bloque 344. El proceso realiza luego un cribado virtual de los aciertos utilizando un sistema de cribado de proteínas virtual, representado en el bloque 346, que es similar al paso 304 en el proceso 300. En algunas realizaciones, el proceso también selecciona una o más secuencias basadas en la energía de enlace, la actividad y la selectividad determinadas por el sistema de selección virtual como se describió anteriormente. El proceso luego evalúa si es necesario realizar una ronda adicional de evolución dirigida de las secuencias seleccionadas en el paso 348. Si es así, el proceso proporciona las secuencias seleccionadas para una ronda adicional de evolución dirigida

in vitro en una nueva iteración, véanse el bloque 342. El cribado y la selección virtuales pueden continuar para las iteraciones, hasta que no sea necesaria una evolución posterior de las secuencias, que puede determinarse mediante criterios preestablecidos. En ese punto, el proceso de diseño de biomoléculas (por ejemplo, enzimas) se termina en el paso 350.

IV. GENERACIÓN DE UNA BIBLIOTECA VARIANTE DE PROTEÍNA

[0183] Las bibliotecas de variantes de proteínas comprenden grupos de proteínas múltiples que tienen uno o más residuos que varían de miembro a miembro en una biblioteca. Estas bibliotecas se pueden generar usando los métodos descritos en este documento y/o cualquier medio adecuado conocido en la técnica. En diversas realizaciones, estas bibliotecas proporcionan enzimas candidatas para el sistema virtual de selección de proteínas. En algunas realizaciones, las bibliotecas pueden proporcionarse y seleccionarse *in silico* en rondas iniciales, y las proteínas resultantes seleccionadas mediante el sistema de selección virtual de una ronda posterior o final pueden secuenciarse y/o rastrearse *in vitro.* Debido a que las rondas iniciales de detección se realizan *en silico,* el tiempo y el costo del cribado pueden reducirse significativamente. El número de proteínas incluidas en una biblioteca de variantes de proteínas puede aumentarse fácilmente en las rondas iniciales de selección en algunas implementaciones en comparación con el examen físico convencional. No se pretende que la presente divulgación se limite a ningún número particular de proteínas en las bibliotecas de proteínas usadas en los métodos de la presente descripción. Además, no se pretende que la presente divulgación se limite a ninguna biblioteca o bibliotecas de variantes de proteínas particulares.

[0184] En un ejemplo, la biblioteca de variantes de proteínas se genera a partir de una o más proteínas de origen natural, que pueden estar codificadas por una familia de genes única en algunas realizaciones, o un panel de enzimas en otras realizaciones. Otros puntos de partida incluyen, pero no se limitan a, recombinantes de proteínas conocidas y/o nuevas proteínas sintéticas. A partir de estas proteínas "semilla" o "inicial", la biblioteca puede generarse mediante diversas técnicas. En un caso, la biblioteca se genera mediante procesos virtuales que reflejan técnicas biológicas o químicas, por ejemplo, recombinación mediada por fragmentación de ADN como se describe en Stemmer (1994) Proceedings of the National Academy of Sciences, USA, 10747-10751 y WO 95/22625., recombinación mediada por oligonucleótidos sintéticos como se describe en Ness et al. (2002) Nature Biotechnology 20: 1251-1255 y WO 00/42561, o ácidos nucleicos que codifican parte o la totalidad de una o más proteínas parentales. Se pueden usar combinaciones de estos métodos (por ejemplo, recombinación de fragmentos de ADN y oligonucleótidos sintéticos) así como otros métodos basados en la recombinación conocidos en la técnica, por ejemplo, WO97/20078 y WO98/27230. Cualquier método adecuado usado para generar bibliotecas de variantes de proteínas encuentra uso en la presente divulgación. De hecho, no se pretende que la presente divulgación se limite a ningún método particular para producir bibliotecas de variantes.

[0185] En algunas realizaciones, una única secuencia de "inicio" (que puede ser una secuencia "antecesora") puede emplearse para definir un grupo de mutaciones usadas en el proceso de modelado. En algunas realizaciones, hay más de una secuencia de inicio. En algunas realizaciones adicionales, al menos una de las secuencias de inicio es una secuencia de tipo silvestre. En ciertas realizaciones, las mutaciones se identifican (a) en la literatura por afectar a la especificidad, selectividad, estabilidad y/o cualquier otra propiedad de interés del sustrato y/o (b) se predicen computacionalmente para mejorar los patrones de plegamiento de proteínas (por ejemplo, empaquetar el interior) residuos de una proteína), mejorar la unión del ligando, mejorar las interacciones de la subunidad, o mejorar los métodos de mezcla familiar entre múltiples homólogos diversos, etc. No se pretende que la presente invención se limite a ninguna elección específica de propiedad(es) de interés o función(es).

[0186] En algunas realizaciones, las mutaciones pueden ser virtualmente introducidas en la secuencia de inicio y las proteínas pueden ser rastreadas virtualmente por sus propiedades beneficiosas. La mutagénesis dirigida al sitio es un ejemplo de una técnica útil para introducir mutaciones, aunque cualquier método adecuado encuentra uso. De este modo, alternativamente o además, los mutantes pueden proporcionarse mediante síntesis génica, saturación de mutagénesis aleatoria, bibliotecas combinatorias semisintéticas de residuos, evolución dirigida, recombinación de secuencia recursiva ("RSR") (véanse, por ejemplo, la Solicitud de Patente de los Estados Unidos Nº de publicación 2006/0223143), barajado de genes, PCR propensa a errores, y/o cualquier otro método adecuado. Un ejemplo de un procedimiento de mutagénesis de saturación adecuado se describe en la Solicitud de Patente de los Estados Unidos Publ. Nº 2010/0093560.

[0187] La secuencia de inicio no necesita ser idéntica a la secuencia de aminoácidos de una proteína de tipo salvaje. Sin embargo, en algunas realizaciones, la secuencia de inicio es la secuencia de una proteína de tipo salvaje. En algunas realizaciones, la secuencia de inicio incluye mutaciones no presentes en la proteína de tipo salvaje. En algunas realizaciones, la secuencia de inicio es una secuencia de consenso derivada de un grupo de proteínas que tiene una propiedad común, por ejemplo, una familia de proteínas.

[0188] En algunas realizaciones, las transformaciones químicas catalizadas que pueden cribarse usando el sistema de selección virtual incluyen, pero no están limitadas a, por ejemplo, reducción de cetona, transaminación, oxidación, hidrólisis de nitrilo, reducción de iminas, reducción de enonas, hidrólisis de acilo y deshalogenación de halohidrina. Los ejemplos de clases de enzimas que pueden proporcionar las múltiples variantes evaluadas incluyen, pero no se

limitan a, reductasas de cetonas, transaminasas, citocromo P450, monooxigenasas de Baeyer-Villiger, oxidasas de monoamina, nitrilasas, reductasas de imina, reductasas de enona, acilasas y deshalogenasas de halohidrina..

[0189] Una lista representativa no limitante de familias o clases de enzimas que pueden servir como fuentes de secuencias parentales incluye, pero no se limita a, las siguientes: oxidorreductasas (EC1); transferasas (EC2); hidroliasas (EC3); liasas (EC4); isomerasas (EC5) y ligasas (EC 6). Subgrupos de oxidoreductasas más específicas pero no limitantes incluyen deshidrogenasas (por ejemplo, deshidrogenasas de alcohol (reductasas de carbonilo), reductasas de xilulosa, reductasas de aldehído, deshidrogenasa de farnesol, deshidrogenasas de lactato, deshidrogenasas de arabinosa, deshidrogenasa de glucosa, deshidrogenasas de fructosa, reductasas de xilosa y deshidrogenasas de succinato), oxidasas (por ejemplo, oxidasas de glucosa, oxidasas de hexosa, oxidasas de galactosa y lacasas), oxidasas de monoamino, lipoxigenasas, peroxidasas, deshidrogenasas de aldehído, reductasas, reductasas de acilo-[acilo-portador-proteína] de cadena larga, deshidrogenasas de acilo-CoA, ene-reductasas, sintasas (por ejemplo, sintasas de glutamato), reductasas de nitrato, mono y di-oxigenasas y catalasas. Los subgrupos de transferasas más específicas pero no limitativas incluyen metilo, amidino y carboxiltransferasas, transquetolasas, transaldolasas, aciltransferasas, glicosiltransferasas, transaminasas, transglutaminasas y polimerasas. Subgrupos de hidrolasas más específicas pero no limitativas incluyen hidrolasas de éster, peptidasas, glicosilasas, amilasas, celulasas, hemicelulosa, xilanasas, quitinasas, glucosidasas, glucanasas, glucoamilasas, acilasas, galactosidasas, pululanasas, fitasas, lactasas, arabinosidasas, nucleosidasas, nitrilasa, fosfatasas, lipasas, fosfolipasas, proteasas, ATPasas y deshalogenasas. Los subgrupos de liasas más específicos pero no limitantes incluyen descarboxilasas, aldolasas, hidratasas, deshidratasa (por ejemplo, anhidrasas carbónicas), sintasas (por ejemplo, sintasas de isopreno, pineno y farneseno), pectinasas (por ejemplo, pectinasas) y deshidrogenasas de halohidrina. Los subgrupos de isomerasas más específicos, pero no limitantes, incluyen racemasas, epimerasas, isomerasas (por ejemplo, isomerasas de xilosa, arabinosa, ribosa, glucosa, galactosa y manosa), tautomerasas y mutasas (por ejemplo, mutasas, fosfomutasas y aminomutasas que transfieren acilo) pero los subgrupos no limitantes de ligasas incluyen sintasas de ésteres. Otras familias o clases de enzimas que pueden usarse como fuentes de secuencias parentales incluyen transaminasas, proteasas, quinasas y sintasas. Esta lista, aunque ilustra ciertos aspectos específicos de las posibles enzimas del divulgación, no se considera exhaustiva y no retrata las limitaciones ni circunscribe el alcance de la divulgación.

[0190] En algunos casos, las enzimas candidatas útiles en los métodos descritos en este documento son capaces de catalizar una reacción enantioselectiva tal como una reacción de reducción enantioselectiva, por ejemplo. Tales enzimas pueden usarse para hacer intermedios útiles en la síntesis de compuestos farmacéuticos, por ejemplo.

[0191] En algunas realizaciones, las enzimas candidatas se seleccionan de endoxilanasas (EC 3.2.1.8); β-oxilosidasas (EC 3.2.1.37); alfa-L-arabinofuranosidasas (EC 3.2.1.55); alfa-glucuronidasas (EC 3.2.1.139); acetilxilanosterasas (EC 3.1.1.72); esterasas de feruloílo (EC 3.1.1.73); esterasas de cumaroílo (EC 3.1.1.73); alfa-galactosidasas (EC 3.2.1.22); beta-galactosidasas (EC 3.2.1.23); beta-mananasas (EC 3.2.1.78); beta-manosidasas (EC 3.2.1.25); endo-poligalacturonasas (EC 3.2.1.15); esterasas metílicas de pectina (EC 3.1.1.11); endo-galactanasas (EC 3.2.1.89); esterasas acetílicas de pectina (EC 3.1.1.6); endopectinasasas (EC 4.2.2.10); liasas de pectato (EC 4.2.2.2); alfa-ramnosidasas (EC 3.2.1.40); exo-poli-alfa-galacturonosidasa (EC 3.2.1.82); 1,4-alfa-galacturonidasa (EC 3.2.1.67); exopolisalacturonasas (EC 4.2.2.9); endoliasas de ramnogalacturonano EC (4.2.2.B3); ramnogalacturonanoacetilterasas (EC 3.2.1.B11); galacturonohidrolasas de ramnogalacturonano (EC 3.2.1.B11); endo-arabinanasas (EC 3.2.1.99); lacasas (EC 1.10.3.2); peroxidasas dependientes de manganeso (EC 1.10.3.2); amilasas (EC 3.2.1.1), glucoamilasas (EC 3.2.1.3), proteasas, lipasas y peroxidasas de lignina (EC 1.11.1.14). Cualquier combinación de una, dos, tres, cuatro, cinco o más de cinco enzimas encuentra uso en las composiciones de la presente descripción. No se pretende que la presente invención se limite a ningún número particular de enzimas y/o clases de enzimas.

[0192] No se pretende que la presente invención se limite a ningún método particular para generar secuencias variadas sistemáticamente, como encuentra uso cualquier método adecuado. En una o más realizaciones de la divulgación, una única secuencia de inicio se modifica de diversas maneras para generar la biblioteca. En algunas realizaciones, la biblioteca se genera variando sistemáticamente los residuos individuales de la secuencia de inicio. El conjunto de secuencias sistemáticamente variadas de una biblioteca se puede diseñar *a priori* usando métodos de diseño de experimentos (DOE) para definir las secuencias en el conjunto de datos. Una descripción de los métodos DOE se puede encontrar en Diamond, WJ (2001) Practical Experiment Designs: for Engineers and Scientists, John Wiley & Sons y en "Practical Experimental Design for Engineers and Scientists" por William J Drummond (1981) Van Nostrand Reinhold Co Nueva York, "Statistics for experimenters" George EP Box, William G Hunter y J. Stuart Hunter (1978) John Wiley and Sons, Nueva York, o, por ejemplo, en la World Wide Web en itl.nist.gov/div898/handbook/. Hay varios paquetes computacionales disponibles para realizar las matemáticas relevantes, incluidos Statistics Toolbox (MAT-LAB®), JMP®, STATISTICA® y STAT-EASE® DESIGN EXPERT®. El resultado es un conjunto de secuencias de datos dispersos ortogonalmente variados y sistemáticamente que es adecuado para el cribado mediante el sistema virtual de cribado de proteínas descrito en este documento. Los conjuntos de datos basados en DOE también se pueden generar fácilmente usando Plackett-Burman o Diseños Factoriales Fraccionales, como se conoce en la técnica. Diamond, WJ (2001).

[0193] Debido a que las rondas iniciales de cribado se pueden realizar *in silico* con alta eficacia, algunas

realizaciones pueden usar algunas o todas las secuencias disponibles para proporcionar la biblioteca de variantes de proteínas cuando el número de variantes suele ser demasiado grande para cribar con métodos físicos convencionales. Por ejemplo, para una secuencia con 15 posiciones, teniendo cada una 20 posibles aminoácidos, hay 300 posiciones posibles frente a pares de aminoácidos, y $\sum_{r=1}^{300}\binom{300}{r}$ diferentes secuencias variantes. En algunas implementaciones, una biblioteca puede incluir cientos, miles, decenas de miles, cientos de miles o más variantes de este grupo posible, dependiendo de la potencia informática disponible y las necesidades de la aplicación. No se pretende que la presente divulgación se limite a ningún número particular de variante en las bibliotecas.

V. SECUENCIACIÓN DE LAS VARIANTES DE PROTEÍNA

[0194] En algunas realizaciones, las variantes de proteína física se usan para generar modelos computacionales de sitios activos de las variantes de proteína usadas en selección virtual como se describió anteriormente. En algunas realizaciones, las variantes de proteínas obtenidas a partir de cribado virtual se generan físicamente usando diversos métodos descritos anteriormente. En algunas realizaciones, las variantes de proteínas generadas físicamente se ensayan para su reacción contra uno o más ligandos de interés. En diversas realizaciones, las secuencias de las variantes de proteínas físicas se determinan por métodos de secuenciación de proteínas, algunos de los cuales se describen adicionalmente a continuación.

[0195] 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 la medida 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.

[0196] La reacción de degradación de Edman permite descubrir la composición de aminoácidos ordenada de una proteína. En algunas realizaciones, los secuenciadores Edman automatizados pueden usarse para determinar la secuencia de variantes de proteína. Los secuenciadores Edman automatizados son capaces de 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 los enlaces se vuelvan a formar
--Separar y purificar las cadenas individuales del complejo de proteínas 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.
--Separar y purificar los fragmentos
--Determinar la secuencia de cada fragmento usando la reacción de degradación de Edman
--Repetir los pasos anteriores aplicando un patrón diferente de escisión para proporcionar lecturas adicionales de secuencias de aminoácidos
--Crear la secuencia de la proteína global de las lecturas de secuencia de aminoácidos

[0197] En diversas implementaciones, los péptidos de más de aproximadamente 50-70 aminoácidos deben dividirse en pequeños fragmentos para facilitar la secuenciación mediante reacciones de Edman. La digestión de secuencias más largas puede realizarse mediante endopeptideas tales como tripsina o pepsina, o mediante reactivos químicos tales como bromuro de cianógeno. Diferentes enzimas dan diferentes patrones de escisión, y la superposición entre fragmentos se puede usar para construir una secuencia global.

[0198] Durante la reacción de degradación de Edman, el péptido a secuenciar se adsorbe en una superficie sólida de un sustrato. En algunas realizaciones, un sustrato adecuado es fibra de vidrio recubierta con polibreno, un polímero catiónico. El reactivo de Edman, fenilisotiocianato (PITC), se agrega al péptido adsorbido, junto con una solución de tampón ligeramente básica de trimetilamina. Esta solución de reacción reacciona con el grupo amino del aminoácido N-terminal. El aminoácido terminal puede separarse selectivamente mediante la adición de ácido anhidro. El derivado se isomeriza para dar una feniltiohidantoina sustituida, que puede lavarse e identificarse mediante cromatografía. Entonces el ciclo puede repetirse.

[0199] En algunas realizaciones, la espectrometría de masas se puede usar para determinar una secuencia de aminoácidos determinando las relaciones de masa a carga de los fragmentos de la secuencia de aminoácidos. Se puede determinar el espectro de masas que incluye los picos correspondientes a los fragmentos cargados de forma 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, en comparación con una base de datos de proteínas secuenciadas previamente para determinar las secuencias de los fragmentos. Este proceso se repite luego con una enzima de digestión diferente, y las superposiciones en las secuencias se usan para construir una secuencia de aminoácidos completa.

[0200] Los péptidos a menudo son más fáciles de preparar y analizar para la espectrometría de masas que las proteínas completas. En algunas realizaciones, la ionización por electrospray se usa para administrar los péptidos al espectrómetro. La proteína se digiere mediante 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 pulveriza en el espectrómetro de masas, y 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.

[0201] 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, diversos métodos de secuenciación de próxima generación, pueden usarse para determinar secuencias de ADN o ARN. En algunas implementaciones, una secuencia de proteína se aisla nuevamente sin conocimiento de los nucleótidos que codifican la proteína. En tales implementaciones, se puede determinar primero una secuencia polipeptídica corta 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 puede usarse para aislar el ARNm que codifica la proteína, que luego puede replicarse en una reacción en cadena de la polimerasa para producir una cantidad significativa de ADN, que luego puede secuenciarse usando métodos de secuenciación de ADN. La secuencia de aminoácidos de la proteína se puede deducir a partir de la secuencia de ADN. En la deducción, es necesario tener en cuenta los aminoácidos eliminados después de que el ARNm ha sido traducido.

[0202] En una o más realizaciones, los datos de secuencia de ácido nucleico pueden usarse en diversas etapas en el proceso de evolución dirigida de proteínas. En una o más realizaciones, pueden obtenerse datos de secuencia usando métodos de secuenciación masiva que incluyen, por ejemplo, secuenciación de Sanger o secuenciación de Maxam-Gilbert, que se consideran los primeros métodos de secuenciación de generación. La secuenciación de Sanger, que implica el uso de terminadores de cadena dideoxi 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 bien 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 - 544 (1977). Otro método de secuenciación masiva es la secuenciación por hibridación, en la que la secuencia de una muestra se deduce en base a sus propiedades de hibridación a una pluralidad de secuencias, por ejemplo, en una micromatriz o chip de gen; véase, por ejemplo, Drmanac, et al., Nature Biotechnology 16, 54 - 58 (1998).

[0203] En una o más realizaciones, los datos de secuencia de ácido nucleico se obtienen usando los 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 adecuados de próxima generación incluyen, entre otros, secuenciación en una sola molécula en tiempo real (por ejemplo, Pacific Biosciences of Menlo Park, California), secuenciación de semiconductores de iones (por ejemplo, Ion Torrent of South San Francisco, California), pirólisis (*p. ej.,* 454 de Branford, Connecticut), secuenciación por ligamiento (*p. ej.,* secuenciación SOLiD propiedad de Life Technologies de Carlsbad, California), secuenciación por síntesis y terminador reversible (*p. ej.,* Illumina of San Diego, California), tecnologías de obtención de imágenes de ácido nucleico tales como microscopía electrónica de transmisión y similares.

[0204] En general, los métodos de secuenciación de próxima generación típicamente usan una etapa 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 unen a los cebadores en el cordón, y luego se inmoviliza para una secuencia posterior. EmPCR se usa en los métodos de Marguilis *et al.* (comercializado por 454 Life Sciences, Branford, CT), Shendure y Porreca *et al.* (también conocido como "polony sequencing") y secuenciación SOLiD, (Applied Biosystems Inc., Foster City, CA). Véanse M. Margulies, et al. (2005) "Genome sequencing in microfabricated high-density picolitre reactors" Nature 437: 376-380; J. Shendure, et al. (2005) "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome" Science 309 (5741): 1728-1732. La amplificación clonal in vitro también se puede llevar a cabo mediante "PCR puente", donde los fragmentos se amplifican sobre cebadores unidos a una superficie sólida. Braslavsky *et al.* desarrolló un método de molécula única (comercializado por Helicos Biosciences Corp., Cambridge, MA) que omite este paso de amplificación, fijando directamente moléculas de ADN a una superficie. I. Braslavsky, et al. (2003) "Sequence information can be obtained from single DNA molecules". Proceedings of the National Academy of Sciences of the United States of America 100: 3960-3964.

[0205] Las moléculas de ADN que están físicamente unidas a una superficie se pueden secuenciar en paralelo. En "secuenciación por síntesis", se construye una cadena complementaria basada en la secuencia de una cadena molde usando una polimerasa de ADN, como la secuenciación electroforética de terminación de colorante, los métodos de terminación reversible (comercializados por Illumina, Inc., San Diego, CA y Helicos Biosciences Corp., Cambridge, MA) utilizan versiones reversibles de los terminadores de colorantes, agregan un nucleótido a la vez y detectan la fluorescencia en cada posición en tiempo real, mediante la eliminación repetida del grupo bloqueante

para permitir la polimerización de otro nucleótido. "Pyrosequencing" también usa 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 (comercializados por 454 Life Sciences, Branford, CT). Véanse M. Ronaghi, et al. (1996). "Real-time DNA sequencing using detection of pyrophosphate release" Analytical Biochemistry 242: 84-89.

[0206] Ejemplos específicos de métodos de secuenciación de próxima generación se describen en más detalle 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.

[0207] La secuenciación de molécula única en tiempo real (también conocida como SMRT) es una secuencia de ADN de molécula individual paralelizada mediante tecnología de síntesis desarrollada por Pacific Biosciences. La secuenciación de una sola molécula en tiempo real utiliza la guía de onda de modo cero (ZMW). Se fija una única enzima de polimerasa ADN en la parte inferior de una ZMW con una sola molécula de ADN como molde. El ZMW es una estructura que crea un volumen de observación iluminado que es lo suficientemente pequeño como para observar que solo un nucleótido de ADN (también conocido como una base) se incorpora por la polimerasa de ADN. Cada una de las cuatro bases de ADN está unida a uno de cuatro tintes fluorescentes diferentes. Cuando un nucleótido es incorporado por la polimerasa de ADN, la etiqueta fluorescente se escinde y se difunde fuera del área de observación de la 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 tinte.

[0208] Otra tecnología de secuenciación de molécula 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 de tSMS, una muestra de ADN se escinde en cadenas de aproximadamente 100 a 200 nucleótidos, y se agrega una secuencia poliA al extremo 3' de cada cadena de ADN. Cada cadena se marca mediante la adición de un nucleótido de adenosina marcado fluorescentemente. Las cadenas de ADN se hibridan luego a 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, los moldes pueden ser a una densidad de unos 100 millones de moldes/cm^2. La célula de flujo se carga luego en un instrumento, por ejemplo, el secuenciador HeliScope™, y un láser ilumina la superficie de la célula de flujo, revelando la posición de cada molde. Una cámara CCD puede mapear la posición de los moldes en la superficie de la célula de flujo. La etiqueta fluorescente del molde se escinde y se lava. La reacción de secuenciación comienza por la introducción de una polimerasa de ADN y un nucleótido marcado fluorescentemente. El ácido nucleico oligo-T sirve como cebador. La polimerasa incorpora los nucleótidos marcados al cebador de una manera dirigida al molde. La polimerasa y los nucleótidos no incorporados se eliminan. Los moldes que tienen la incorporación dirigida del nucleótido marcado fluorescentemente se detectan 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 fluorescentemente hasta que se alcanza la longitud de lectura deseada. La información de secuencia se recoge con cada paso de adición de nucleótidos. La secuenciación completa del genoma mediante tecnologías de secuenciación de molécula única excluye o generalmente 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 las copias de esa muestra.

[0209] 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 hebra complementaria basada en la secuencia de una hebra de molde. Un micropocillo que contiene una cadena de molde ADN a secuenciar se inunde 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 homopoliméricas están presentes en la secuencia de molde, se incorporarán múltiples moléculas de dNTP en un solo ciclo. Esto conduce a una cantidad 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 ópticas modificadas. La secuenciación de semiconductores de iones también se puede denominar secuenciación de torrente iónico, secuenciación mediada por pH, secuenciación de silicio o secuenciación de semiconductores.

[0210] En la pirosecuenciación, se hace reaccionar el ion pirofosfato liberado por la reacción de polimerización con adenosina 5' fosfosulfato por ATP sulfurilasa para producir ATP; el ATP luego conduce la conversión de luciferina a oxiluciferina más luz por la luciferasa. Al ser la fluorescencia transitoria, no es necesario un paso separado para eliminar la fluorescencia en este método. Se agrega un tipo de desoxirribonucleótido trifosfato (dNTP), y se discierne la información de la secuencia según la cual el dNTP genera una señal significativa en un sitio de reacción. El instrumento Roche GS FLX disponible en el mercado adquiere secuencia utilizando este método. Esta técnica y sus aplicaciones se discuten en detalle, por ejemplo, en Ronaghi et al., Analytical Biochemistry 242, 84-89 (1996) y Margulies et al., Nature 437, 376-380 (2005) (corrección en Nature 441, 120 (2006)). Una tecnología de pirosecuenciación disponible en el mercado es la secuenciación 454 (Roche) (por ejemplo, como se describe en

Margulies, M. et al., Nature 437: 376-380 [2005]).

[0211] En la secuenciación de la ligación, una enzima ligasa se utiliza para unirse a un oligonucleótido de cadena parcialmente doble con un saliente en el ácido nucleico que se está secuenciado, que tiene un saliente; para que ocurra la ligación, los aleros deben ser complementarios. Las bases en el saliente del oligonucleótido parcialmente bicatenario pueden identificarse de acuerdo con un fluoróforo conjugado con el oligonucleótido parcialmente bicatenario y/o con un oligonucleótido secundario que se 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 unión, tal como mediante una enzima de restricción de tipo II, por ejemplo, Bbvl, que corta en un sitio a una distancia fija de su sitio de reconocimiento (que se incluyó en el oligonucleótido parcialmente de doble cadena). Esta reacción de escisión expone un nuevo saliente justo aguas arriba del saliente 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 ligación 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 molde para la secuenciación de ligación.

[0212] Un ejemplo disponible comercialmente de tecnología de secuenciación de ligación es la tecnología SOLiD™ (Bio-sistemas aplicados). En la secuenciación mediante ligación SOLiD™, el ADN genómico se corta en fragmentos y los adaptadores se unen a los extremos 5' y 3' de los fragmentos para generar una biblioteca de fragmentos. Alternativamente, los adaptadores internos se pueden introducir ligando adaptadores a los extremos 5' y 3' de los fragmentos, circulando los fragmentos, digiriendo el fragmento circularizado para generar un adaptador interno, y uniendo adaptadores a los extremos 5' y 3' de los fragmentos resultantes para generar una biblioteca de par emparejado. A continuación, las poblaciones de perlas clonales se preparan en microrreactores que contienen perlas, cebadores, molde y componentes de PCR. Después de la PCR, los moldes se desnaturalizan y las perlas se enriquecen para separar las perlas con moldes extendidos. Los moldes de 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 ligamiento de oligonucleótidos parcialmente aleatorios con una base determinada central (o un par de bases) que se identifica mediante un fluoróforo específico. Después de registrar un color, el oligonucleótido ligado se escinde y se retira y el proceso se repite a continuación.

[0213] En la secuencia de terminación reversible, un análogo de nucleótido marcado con colorante fluorescente que es un terminador de cadena reversible debido a la presencia de un grupo de bloqueo se incorpora en una reacción de extensión de base única. La identidad de la base se determina de acuerdo con el fluoróforo; en otras palabras, cada base está emparejada con un fluoróforo diferente. Después de que se adquieren los datos de fluorescencia/secuencia, el fluoróforo y el grupo de bloqueo se eliminan químicamente, y el ciclo se repite para adquirir la siguiente base de información de la secuencia. El instrumento Illumina GA funciona con este método. Esta técnica y sus aplicaciones se discuten 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).

[0214] Un ejemplo disponible comercialmente de método de secuenciación del terminador reversible es la síntesis de secuenciación por caso de Illumina y secuenciación reversible a base de terminador (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 planar, ópticamente transparente, sobre la que se unen los anclajes de oligonucleótidos. El ADN molde se repara en el extremo para generar extremos romos fosforilados 5', y la actividad de polimerasa del fragmento Klenow se usa para adición de una única base A al extremo 3' de los fragmentos de ADN fosforados embotados. 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 eficacia de ligación. Los oligonucleótidos adaptadores son complementarios a los anclajes de células de flujo. En condiciones de dilución limitante, se añade ADN de molde monocatenario modificado con adaptador a la célula de flujo y se inmoviliza mediante hibridación con los anclajes. Los fragmentos de ADN unidos se amplían y se amplifican en puente para crear una célula de flujo de secuenciación ultra-alta densidad con cientos de millones de clústeres, cada uno con aproximadamente 1.000 copias del mismo molde. Los moldes se secuencian utilizando una robusta tecnología de secuenciación por síntesis de ADN de cuatro colores que emplea terminadores reversibles con tintes fluorescentes extraíbles. La detección de fluorescencia de alta sensibilidad se logra utilizando la excitación láser y la óptica de reflexión interna total. Las lecturas de secuencias cortas de aproximadamente 20-40 pb, por ejemplo, 36 pb, se alinean contra un genoma de referencia con máscara repetida y el mapeo único de las lecturas cortas de secuencia al genoma de referencia se identifica utilizando un software de canalización de análisis de datos especialmente desarrollado. Los genomas de referencia no enmascarados repetidos también se pueden usar. Si se usan genomas de referencia enmascarados repetidamente o enmascarados no repetidos, solo se contabilizan los mapas exclusivos del genoma de referencia. Después de completar la primera lectura, los moldes pueden regenerarse *in situ* para permitir una segunda lectura desde el extremo opuesto de los fragmentos. Por lo tanto, se puede usar la secuencia final de un solo extremo o apareado de los fragmentos de ADN. Se realiza la secuenciación parcial de los fragmentos de ADN presentes en la muestra, y se cuentan las etiquetas de secuencia que comprenden lecturas de longitud predeterminada, por ejemplo, 36 pb, se mapean a un genoma de referencia conocido.

[0215] En secuenciación de nanoporos, una sola molécula de ácido nucleico de cadena está roscada a través de un poro, por ejemplo, utilizando una fuerza de conducción electroforética, y la secuencia se deduce mediante el análisis de los datos obtenidos como la molécula de ácido nucleico de una sola hebra pasa a través del poro. Los datos pueden ser datos de corriente iónica, en los que cada base altera la corriente, por ejemplo, bloqueando parcialmente la corriente que pasa a través del poro a un grado diferente y distinguible.

[0216] En otra realización ilustrativa, pero no limitativa, los métodos descritos en este documento comprenden obtener información de secuencia usando microscopía electrónica de transmisión (TEM). El método comprende utilizar imágenes de microscopía electrónica de transmisión de resolución de átomo único de ADN de alto peso molecular (150 kb o superior) marcadas selectivamente con marcadores de átomos pesados y disponer estas moléculas en películas ultradelgadas en matrices ultradensas paralelas (3nm de hebra a hebra) con espaciado consistente de base a base. El microscopio electrónico se usa para obtener imágenes de las moléculas en las películas para determinar la posición de los marcadores de átomos pesados y extraer la información de la secuencia de bases del ADN. El método se describe adicionalmente en la publicación de patente PCT WO 2009/046445.

[0217] En otra realización ilustrativa, pero no limitativa, los métodos descritos en este 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 orificios pequeños (~50 nm) como guía de ondas de modo cero (véanse, por ejemplo, Levene et al., Science 299, 682-686 (2003)). La superficie de aluminio está protegida contra la unión de polimerasa de ADN por química de polifosfonato, por ejemplo, química de polivinilfosfonato (véanse, por ejemplo, Korla Ch et al., Proceedings of the National Academy of Sciences of the United States of America 105, 1176-1181 (2008)). Esto da como resultado la unión preferencial de las moléculas de polimerasa de ADN a la sílice expuesta en los orificios del revestimiento de aluminio. Esta configuración permite que los fenómenos de ondas evanescentes se usen para reducir el fondo de fluorescencia, lo que permite el uso de concentraciones más altas de dNTP marcados fluorescentemente. El fluoróforo está unido al fosfato terminal de los dNTP, de modo que la fluorescencia se libera con la incorporación del dNTP, pero el fluoróforo no permanece unido al nucleótido recién incorporado, lo que significa que el complejo está inmediatamente listo para otra ronda de incorporación. Mediante este método, puede detectarse la incorporación de dNTP en un complejo de primer molde individual presente en los orificios del recubrimiento de aluminio. Véase, por ejemplo, Eid et al., Science 323, 133-138 (2009).

VI. ENSAYO DE GENES Y VARIANTES DE PROTEÍNA

[0218] 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 la actividad (o se usan en reacciones de transcripción *in vitro* para fabricar productos que se criban). Además, los ácidos nucleicos que codifican variantes de proteínas se pueden enriquecer, secuenciar, expresar, amplificar *in vitro* o tratar en cualquier otro método recombinante común.

[0219] Los textos generales que describen técnicas de biología molecular útiles en este documento, que incluyen clonación, mutagénesis, construcción de bibliotecas, ensayos de cribado, cultivo celular y similares incluyen Berger y Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volumen 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (2ª ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, Nueva York, 1989 (Sambrook) y Current Protocols in Molecular Biology, FM Ausubel y otros, 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, que incluyen células de plantas y animales, con ácidos nucleicos están generalmente disponibles, como lo son los métodos para expresar proteínas codificadas por tales ácidos nucleicos. Además de Berger, Ausubel y Sambrook, las referencias generales útiles para el cultivo de células animales incluyen Freshney (Culture of Animal Cells, un Manual of Basic Technique, tercera edición Wiley-Liss, Nueva York (1994)) y las referencias citadas allí, 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 clonación de células vegetales, cultivo y regeneración 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) Plant Cell, Tissue and Organ Culture; Fundamental Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg, Nueva York) (Gamborg). Una variedad de medios de cultivo celular se describen en Atlas and 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 Catalogue (1998) de Sigma-Aldrich, Inc (St Louis, MO) (Sigma-LSRCCC) y, por ejemplo, el Plant Culture Catalogue and supplement (1997) también de Sigma-Aldrich, Inc (St Louis, MO) (Sigma-PCCS).

[0220] 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, incluyendo reacciones en cadena de la polimerasa (PCR), reacciones de cadena de ligasa (LCR), amplificaciones de Qβ-replicasa y otras técnicas mediadas por polimerasa ARN (p. ej., NASBA). Estas técnicas se encuentran en Berger, Sambrook y Ausubel, *supra,* así como en Mullis et al., (1987), Patente de los Estados Unidos 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 et al. (1989) Proc. Natl. Acad. Sci. EE.UU. 86, 1173; Guatelli et al. (1990) Proc. Natl. Acad. Sci. EE.UU. 87, 1874; Lomell et al. (1989) J. Clin. Chem 35, 1826; Landegren et al., (1988) Science 241, 1077 - 1080; Van Brunt (1990) Biotechnology 8, 291 - 294; Wu y Wallace, (1989) Gene 4, 560; Barringer et al. (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 de amplificación de ácidos nucleicos grandes por PCR se resumen en Cheng et al. (1994) Nature 369: 684-685 y las referencias en las mismas, en las que se generan amplicones de PCR de hasta 40 kb. Un experto apreciará que, esencialmente, cualquier ARN se puede convertir en un ADN bicatenario adecuado para la digestión de restricción, la expansión de PCR y la secuenciación usando transcriptasa inversa y una polimerasa. Véanse, Ausubel, Sambrook y Berger, *todos supra.*

[0221] En un método preferido, las secuencias reensambladas se verifican para la incorporación de oligonucleótidos de recombinación basados en la familia. Esto se puede hacer 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 por PCR adicionales también son particularmente útiles. Por ejemplo, se ha realizado la secuenciación directa de amplicones generados por PCR al incorporar selectivamente nucleótidos resistentes a nucleasas embrionarias en los amplicones durante la PCR y la digestión de los amplicones con una nucleasa para producir fragmentos de molde de tamaño (Porter et al., 1997) Nucleic Acids Research 25 (8): 1611 - 1617). En los métodos, se realizan cuatro reacciones de PCR en un molde, 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 de forma estocástica en productos de PCR en posiciones variables a lo largo del amplicón de PCR en un conjunto anidado de fragmentos de PCR del molde. Una exonucleasa que está bloqueada por nucleótidos embebidos incorporados se usa para escindir los amplicones de PCR. Los amplicones escindidos se separan luego por tamaño usando electroforesis en gel de poliacrilamida, proporcionando la secuencia del amplicón. Una ventaja de este método es que utiliza menos manipulaciones bioquímicas que la realización de la secuenciación de estilo estándar de Sanger de amplicones de PCR.

[0222] Los genes sintéticos son susceptibles de clonación convencional y enfoques de expresión; por lo tanto, las propiedades de los genes y las proteínas que codifican pueden examinarse fácilmente después de su expresión en una célula huésped. Los genes sintéticos también pueden usarse para generar productos polipeptídicos mediante transcripción y traducción *in vitro* (libre de células). Por lo tanto, los polinucleótidos y polipéptidos pueden examinarse para determinar su capacidad de unirse a una variedad de ligandos, moléculas pequeñas e iones predeterminados, o sustancias poliméricas y heteropoliméricas, que incluyen otras proteínas y epítopos polipeptídicos, así como también paredes de células microbianas, partículas virales, superficies y membranas.

[0223] Por ejemplo, se pueden usar muchos métodos físicos para detectar polinucleótidos que codifican fenotipos asociados con catálisis de reacciones químicas por cualquiera de los polinucleótidos directamente, o por polipéptidos codificados. Únicamente con fines de ilustración, y dependiendo de las características específicas de reacciones químicas particulares determinadas de interés, estos métodos pueden incluir una multitud de técnicas conocidas en la técnica que representan una diferencia física entre sustrato(s) y producto(s), o por 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 infrarrojo (calor)). Estos métodos también pueden seleccionarse de cualquier combinación de los siguientes: espectrometría de masas; resonancia magnética nuclear; materiales marcados isotópicamente, particiones y métodos espectrales que representan la distribución de isótopos o la formación de productos etiquetados; métodos espectrales y químicos para detectar los cambios que acompañan a las composiciones iónicas o elementales de los productos de reacción (incluidos los cambios en el pH, iones orgánicos e inorgánicos y similares). Otros métodos de ensayos físicos, adecuados para uso en los métodos de la presente memoria, pueden basarse en el uso de biosensores específicos para producto(s) de reacción, incluidos aquellos que comprenden anticuerpos con propiedades indicadoras, o aquellos basados en reconocimiento de afinidad in vivo junto con expresión y actividad de un gen informador. Los ensayos acoplados a enzimas para la detección de productos de reacción y las selecciones de vida celular-muerte-crecimiento *in vivo* también pueden usarse cuando sea apropiado. Independientemente de la naturaleza específica de los ensayos físicos, todos se usan para seleccionar una actividad deseada, o una combinación de actividades deseadas, provistas o codificadas por una biomolécula de interés.

[0224] 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 unión a componentes inmovilizados, viabilidad celular u organismal, producción de composiciones informadoras y similares.

[0225] Los 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 van a observar los efectos del tiempo de concentración o incubación, cada 5-10 pocillos puede

analizar una única variante (por ejemplo, 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 1.536 pocillos, entonces una sola placa puede analizar fácilmente de aproximadamente 100 a aproximadamente 1.500 reacciones diferentes. Es posible analizar varias placas diferentes por día; Las pantallas de ensayo para hasta aproximadamente 6.000-20.000 ensayos diferentes (es decir, que implican diferentes ácidos nucleicos, proteínas codificadas, concentraciones, etc.) son posibles 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 puede proporcionar métodos de ensayo de microfluidos de muy alto rendimiento.

[0226] Los sistemas de selección de alto rendimiento están disponibles comercialmente (véanse, por ejemplo, Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, etc.). Estos sistemas típicamente automatizan procedimientos completos que incluyen todas las muestras y pipeteo de reactivos, dispensación de líquidos, incubaciones programadas y lecturas finales de la microplaca en detector(es) apropiado(s) para el ensayo. Estos sistemas configurables proporcionan un alto rendimiento y un inicio rápido, así como un alto grado de flexibilidad y personalización.

[0227] Los fabricantes de tales sistemas proporcionan protocolos detallados para varios ensayos de selección de alto rendimiento. De este modo, por ejemplo, Zymark Corp. proporciona boletines técnicos que describen sistemas de exploración para detectar la modulación de la transcripción de genes, unión de ligandos, y similares.

[0228] 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 PC (Intel x86 o MAC OS compatible con chips pentium, familia WINDOWS™, o Equipos basados en UNIX (por ejemplo, estación de trabajo SUN™).

[0229] Los sistemas para el análisis generalmente incluyen una computadora digital específicamente programada para realizar algoritmos especializados que usan software para dirigir uno o más pasos de uno o más de los métodos de este documento, y, opcionalmente, también incluyen, por ejemplo, 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, 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 transferencia de líquidos de alto rendimiento por parte del armadura robótica de control de líquido y, opcionalmente, un escáner de imagen para digitalizar las señales de etiqueta de los 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, el software de interpretación de datos interpreta la medición de la intensidad de la etiqueta de la sonda para mostrar si la sonda marcada se hibrida con el ADN en el soporte sólido.

[0230] 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. Utilizando un selector automático de colonias (por ejemplo, Q-bot, Genetix, Reino Unido), se identifican colonias o placas, se recogen y hasta 10.000 mutantes diferentes inoculados en placas de microtitulación de 96 pocillos que contienen dos bolas de vidrio de 3 mm/pocillo. El Q-bot no selecciona una colonia completa, sino que inserta un alfiler en el 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 pasador está en la colonia, el número de inmersiones para inocular el medio de cultivo, y el tiempo en que el pasador está en ese medio, cada tamaño de inóculo de efecto, y cada parámetro se puede controlar y optimizar.

[0231] El proceso uniforme de recogida de colonia automatizada como la Q-bot disminuye el error de manipulación humana y aumenta la velocidad de establecimiento de cultivos (aproximadamente 10.000/4 horas). Estos cultivos se agitan opcionalmente en una incubadora con control de temperatura y humedad. 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 se pueden aislar mediante dilución limitante. Como también se ha descrito anteriormente, las placas o células que constituyen bibliotecas también se pueden cribar directamente para la producción de proteínas, ya sea detectando la hibridación, la actividad de proteínas, la unión de proteínas a anticuerpos o similares. Para aumentar las posibilidades de identificar un grupo de tamaño suficiente, se puede usar un precribado que aumenta el número de mutantes procesados por 10 veces. El objetivo del cribado principal consiste en identificar rápidamente mutantes que tengan títulos de producto iguales o mejores que las cepas parentales y mover estos mutantes solo hacia adelante para el cultivo de células líquidas para su posterior análisis.

[0232] Un enfoque para seleccionar bibliotecas diversas consiste en usar un procedimiento de fase sólida masivamente paralelo para seleccionar células que expresan variantes de polinucleótidos, por ejemplo, polinucleótidos que codifican variantes de enzimas. Están disponibles aparatos de cribado de fase sólida

masivamente paralelos que utilizan absorción, fluorescencia o FRET. Véase, por ejemplo, la patente de EE.UU. Nº 5.914.245 de Bylina, et al. (1999); véanse también, http://www|.|kairos-scientific.com/; Youvan et al. (1999) " Fluorescence Imaging Micro-Spectrophotometer (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; y Youvan et al. (1999) "Calibration of Fluorescence Resonance Energy Transfer in Microscopy Using Genetically Engineered GFP Derivatives on Nickel Chelating Beads" publicado en www|.|Kairos-scientific.com. Después del escrutinio 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 a continuación como se establece en este documento para diseñar una nueva biblioteca de variantes de proteínas.

[0233] Del mismo modo, un número de sistemas robóticos bien conocidos también se han desarrollado para la química en fase de disolución 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, Massachusetts, Orca, Beckman Coulter, Inc. (Fullerton, CA)) que imitan las operaciones sintéticas manuales 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 desarrolladas como se describe en este documento. La naturaleza y la implementación de las modificaciones a estos dispositivos (si existen) para que puedan funcionar como se describe en este documento serán evidentes para las personas expertas en la técnica relevante.

VII. APARATOS Y SISTEMAS DIGITALES

[0234] Como debería ser evidente, las realizaciones descritas en este documento emplean procesos que actúan bajo el control de instrucciones y/o datos almacenados en o transferidos a través de uno o más sistemas informáticos. Las realizaciones descritas en este documento también se refieren a sistemas y aparatos (por ejemplo, equipos) para realizar estas operaciones. En algunas realizaciones, el aparato está especialmente diseñado y/o construido para los fines requeridos, o puede ser una computadora de propósito general activada selectivamente o reconfigurada por un programa informático y/o una estructura de datos almacenada en la computadora. Los procesos proporcionados por la presente divulgación no están intrínsecamente relacionados con ninguna computadora particular u otro aparato específico. En particular, diversas 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. Una forma de realización de una estructura particular para una variedad de estas máquinas se describe a continuación.

[0235] Además, ciertas realizaciones de la presente divulgación se refieren a medios legibles por computadora o productos de programas informáticos que incluyen instrucciones y/o datos de programas (que incluyen estructuras de datos) para realizar diversas operaciones implementadas por computadora. Los ejemplos de medios legibles por computadora incluyen, pero no se limitan a, medios magnéticos tales como discos duros; medios ópticos tales como dispositivos de CD-ROM y dispositivos holográficos; medios magneto-ópticos; y dispositivos de memoria semiconductores, como memoria flash. Los dispositivos de hardware tales como los dispositivos de memoria de solo lectura (ROM) y los dispositivos de memoria de acceso aleatorio (RAM) se pueden configurar para almacenar las instrucciones del programa. Los dispositivos de hardware tales como los circuitos integrados específicos de la aplicación (ASIC) y los dispositivos lógicos programables (PLD) pueden configurarse para ejecutar y almacenar las instrucciones del programa. No se pretende que la presente divulgación se limite a ningún medio en particular legible por computadora o cualquier otro producto de programa informático que incluya instrucciones y/o datos para realizar operaciones implementadas por computadora.

[0236] Ejemplos de instrucciones de programa incluyen, pero no se limitan a, código de bajo nivel tal como el producido por un compilador, y archivos que contienen un código de nivel superior que puede ser ejecutado por la computadora usando un intérprete. Además, las instrucciones del programa incluyen, entre otras, código de máquina, código fuente y cualquier otro código que controle directa o indirectamente el funcionamiento de una máquina informática de acuerdo con la presente descripción. El código puede especificar entrada, salida, cálculos, condicionales, ramas, bucles iterativos, etc.

[0237] En un ejemplo ilustrativo, los métodos de incorporación de código 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 cargan en un dispositivo informático configurado apropiadamente hacen que el dispositivo realice una detección virtual de una o más variantes de biomoléculas que interactúan con uno o más ligandos. La Figura 4 muestra un ejemplo de dispositivo digital 800 que es un aparato lógico que puede leer instrucciones desde el medio 817, el puerto de red 819, el teclado de entrada de usuario 809, la entrada de usuario 811 u otros medios de entrada. El aparato 800 puede utilizar después esas instrucciones para dirigir operaciones estadísticas en el espacio de datos, por ejemplo, para evaluar una relación geométrica entre un resto ligando y una o más características de un sitio activo, cofactor, etc. (por ejemplo, para determinar una distancia entre la posición de un sustrato nativo en un sitio activo y la posición de un sustrato considerado en el sitio activo de una variante de proteína). Un tipo de aparato lógico que puede incorporar realizaciones divulgadas es un sistema informático como en el sistema informático 800

que comprende la CPU 807, el teclado 809 de dispositivos de entrada de usuario opcional y el dispositivo señalador GUI 811, así como componentes periféricos tales como unidades de disco 815 y pantalla 805 (que muestra cadenas de caracteres GO modificadas y proporciona una selección simplificada de subconjuntos de tales cadenas de caracteres por un usuario. El medio fijo 817 se utiliza opcionalmente para programar el sistema general y puede incluir, por ejemplo, un medio óptico o magnético de disco u otro elemento de almacenamiento de memoria electrónica. El puerto de comunicación 819 se puede usar para programar el sistema y puede representar cualquier tipo de conexión de comunicación.

[0238] Ciertas realizaciones también pueden incorporarse dentro de la circuitería de un circuito integrado específico de la aplicación (ASIC) o un dispositivo lógico programable (PLD). En tal caso, las realizaciones se implementan en un lenguaje descriptivo legible por computadora que se puede usar para crear un ASIC o PLD. Algunas realizaciones de la presente divulgación se implementan dentro del circuito o procesadores lógicos de una variedad de otros aparatos digitales, tales como PDA, sistemas de computadora portátil, pantallas, equipos de edición de imágenes, etc.

[0239] En algunas realizaciones, la presente divulgación se refiere a un producto de programa informático que comprende uno o más medios de almacenamiento legibles por ordenador que tienen almacenadas en él instrucciones ejecutables por computadora que, cuando son ejecutadas por uno o más procesadores de un sistema informático, dan lugar a que el sistema informático implemente un método para el cribado virtual de variantes de proteínas y/o la evolución dirigida *in silico* de proteínas que tienen actividad deseada. Tal método puede ser cualquier método descrito en la presente memoria, tal 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 selecciona enzimas que tienen actividad catalítica deseada y selectividad. En algunas realizaciones, el método puede desarrollar adicionalmente bibliotecas variantes a partir de variantes que han sido altamente clasificadas por el proceso de selección. Las bibliotecas variantes se pueden usar en la evolución y cribado dirigidos reiteradamente, que pueden dar como resultado enzimas de propiedades beneficiosas deseadas.

[0240] 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 activos de una pluralidad de variantes como se describe en este documento. En diversas realizaciones, los métodos para determinar el acoplamiento implican evaluar la energía de enlace entre una posición del sustrato y la enzima. Para una variante de proteína que se acopla satisfactoriamente con el ligando, el sistema de selección de proteína virtual considera una pluralidad de posiciones de la representación computacional del ligando en el sitio activo de la variante proteínica en consideración, y determina cuál de las posiciones es activa. En diversas realizaciones, los métodos para determinar las posiciones activas implican evaluar las restricciones geográficas que definen un rango de posiciones relativas de uno o más átomos en el ligando y uno o más átomos en la proteína y/o cofactor asociados con la proteína.

VIII. MODALIDADES EN SITIOS WEB Y COMPUTACIÓN EN LA NUBE

[0241] 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 el correo electrónico, ftp, la World Wide Web ("WWW") y otros servicios, incluidos los servicios de seguridad. Se puede entender que el servicio WWW permite que un sistema informático servidor (por ejemplo, un servidor web o un sitio web) envíe páginas web de información a un dispositivo remoto de información del cliente o sistema informático. El sistema informático del cliente remoto puede mostrar las páginas web. Generalmente, 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 de computadora 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 de computadora del cliente recibe esa página web, puede mostrar la página web usando un navegador o puede interactuar con la página web o la interfaz de otra manera. Un navegador es un módulo lógico que efectúa la solicitud de páginas web y muestra o interactúa con páginas web.

[0242] Actualmente, las páginas web que se pueden visualizar se definen típicamente usando un Lenguaje de Marcado de Hipertexto ("HTML"). HTML proporciona un conjunto estándar de etiquetas que definen cómo 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 servidor u otros sistemas informáticos servidores. Las URL también pueden indicar otros tipos de interfaces, incluidos los scripts CGI o las interfaces ejecutables, que los dispositivos de información utilizan para comunicarse con dispositivos o servidores de información remota sin mostrar necesariamente la información a un usuario.

[0243] Internet es especialmente propicio para proporcionar servicios de información a uno o más clientes remotos. Los servicios pueden incluir elementos (p. ej., música o cotizaciones bursátiles) que se envían 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 ser administrados a través de canales de distribución convencionales (por ejemplo, un proveedor común). Los servicios también pueden incluir el manejo de pedidos de artículos, tales como reservas de aerolíneas o teatros, a los que un comprador accede en un momento posterior. Un sistema de computadora de servidor puede proporcionar una versión electrónica de una interfaz que enumera elementos o servicios que están disponibles. Un usuario o un posible comprador puede acceder a la interfaz mediante un navegador y seleccionar varios elementos de interés. Cuando el usuario haya 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 orden 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 número de cuenta) o información adicional necesaria para completar el servicio, como información de vuelo.

[0244] Entre los servicios de particular interés que se pueden proporcionar a través de Internet y sobre otras redes se encuentran datos biológicos y bases de datos biológicos. Dichos servicios incluyen una variedad de servicios provistos 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 el conocimiento sobre biología molecular, bioquímica y genética; facilitar el uso de tales bases de datos y software por parte de la comunidad médica y de investigación; coordinar los esfuerzos para recopilar información sobre biotecnología tanto a nivel nacional como internacional; y realizar investigaciones sobre métodos avanzados de procesamiento de información basado en computadora para analizar la estructura y función de moléculas biológicamente importantes.

[0245] 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 de Biología Molecular Europeo (EMBL) y la Base de Datos de ADN de Japón (DDBJ), e incluye datos de secuencia de patentes y la Oficina de Marcas. Además de GenBank®, NCBI apoya y distribuye una variedad de bases de datos para las comunidades médicas y científicas. Incluyen la herencia mendeliana en línea en el hombre (OMIM), la base de datos de modelado molecular (MMDB) de las estructuras de proteínas 3D, la colección de secuencias genéticas humanas únicas (UniGene), un mapa genético del genoma humano, el navegador taxonómico y el proyecto de anatomía de genoma de cancer (CGAP), en colaboración con el National Cancer Institute. Entrez es el sistema de búsqueda y recuperación de NCBI que proporciona 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 en este documento, es un programa de búsqueda de similitud de secuencia desarrollado en NCBI para identificar genes y características genéticas que pueden ejecutar búsquedas de secuencias contra la base de datos de ADN completa. Las herramientas de software adicionales proporcionadas por NCBI incluyen: Buscador de marco de lectura abierto (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 la WWW o en FTP o en servidores de correo electrónico. Más información está disponible en www|.|Ncbi.nlm.nih.gov.

[0246] Algunos datos biológicos disponibles a través de Internet son datos que generalmente se ven con un "complemento" de navegador especial u otro código ejecutable. Un ejemplo de este tipo de sistema es CHIME, un complemento de navegador que permite una visualización interactiva tridimensional de estructuras moleculares, incluidas las estructuras biológicas moleculares. Se puede encontrar más información sobre CHIME en www|.|mdlchime.com/chime/.

[0247] Varias compañías e instituciones proporcionan sistemas en línea para ordenar compuestos biológicos. Ejemplos de tales sistemas se pueden encontrar en www|.|Genosys.com/oligo_custinfo.cfm o www|.|Genomictechnologies.com/Qbrowser2_FP.html. Típicamente, estos sistemas aceptan algún descriptor de un compuesto biológico deseado (tal como un oligonucleótido, cadena de ADN, cadena de ARN, secuencia de aminoácidos, etc.) y luego el compuesto solicitado se fabrica y se envía al cliente en una solución líquida u otra forma apropiada.

[0248] Ya que los métodos proporcionados en el presente documento pueden implementarse en un sitio web como se describe adicionalmente a continuación, los resultados de cálculo o resultados físicos que implican polipéptidos o polinucleótidos producidos por algunas realizaciones de la divulgación se pueden proporcionar a través de Internet de una manera similar a la información biológica y compuestos descritos anteriormente.

[0249] Para ilustrar adicionalmente, los métodos de esta invención se pueden implementar 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 generalizada o local dedicada o de área amplia y, en ciertas realizaciones preferidas, las

computadoras pueden ser componentes de Intranet o Internet.

[0250] En una realización de Internet, un sistema cliente típicamente ejecuta un navegador web y está acoplado a una computadora de 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 suele ser, pero no necesariamente, un programa como HTTP Daemon de IBM u otro daemon de www (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 a través de un sistema inalámbrico. A su vez, la computadora de 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.

[0251] Como se mencionó, un usuario de un cliente conectado a Intranet o Internet puede hacer que el cliente solicite recursos que son parte del sitio o sitios web que aloja(n) la(s) aplicación(es) proporcionando una implementación de los métodos de esta invención. Los programas del servidor procesan la solicitud para devolver los recursos especificados (suponiendo que estén disponibles actualmente). La convención de nomenclatura estándar (es decir, el localizador uniforme de recursos ("URL")) abarca varios tipos de nombres de ubicación, actualmente incluye subclases como el protocolo de transporte de hipertexto ("http"), el protocolo de transporte de archivos ("ftp"), el gopher y el 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 de la existencia de nuevos recursos que no había solicitado específicamente.

[0252] El software que implementa el (los) método(s) de esta invención puede ejecutarse localmente en el servidor que aloja el sitio web en una verdadera arquitectura cliente-servidor. Por lo tanto, la computadora del cliente envía las solicitudes al servidor que ejecuta los procesos solicitados localmente y luego descarga los resultados nuevamente al cliente. Alternativamente, los métodos de esta invención se pueden implementar en un formato de "múltiples niveles" en el que un componente del (de los) método(s) se realiza(n) localmente por el cliente. Esto puede implementarse mediante software descargado del servidor a petición del cliente (por ejemplo, una aplicación Java) o puede implementarse mediante un software "permanentemente" instalado en el cliente.

[0253] En una realización, la(s) aplicación(es) que implementa(n) los métodos de esta invención se dividen en marcos. En este paradigma, es útil ver una aplicación no tanto como una colección de funciones o funcionalidades sino, en cambio, 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 funcionalidad. 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 subaplicación). Así, por ejemplo, uno o más cuadros pueden proporcionar funcionalidad para introducir 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.

[0254] En ciertas realizaciones, los métodos de esta invención se implementan como uno o más marcos que proporcionan, por ejemplo, las siguientes funcionalidades: función(es) 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 producto de la misma longitud que una o más de las cadenas de caracteres iniciales; funciones para agregar (colocar) las cadenas de producto a una colección de cadenas; funciones para crear y manipular representación computacional/modelos 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 la dinámica molecular a modelos moleculares; funciones para calcular diversas restricciones entre las moléculas que afectan las reacciones químicas que implican las moléculas (por ejemplo, la distancia o el ángulo entre un resto del sustrato y un sitio activo de la enzima); y funciones para implementar cualquier característica establecida aquí.

[0255] 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 representaciones 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 a una función de codificación y/o de generación de diversidad.

[0256] Los métodos para implementar realizaciones de intranet y/o intranet de procesos de acceso informático y/o informático son bien conocidos por los expertos en la técnica y están documentados con gran detalle (véanse, 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, 2-5 de junio de 1992, SIGMOD Record, volumen 21, edición 2, junio de 1992; Stonebraker, M., Editor; ACM Press, pp. 383-392; ISO-ANSI, Working Draft, "Information Technology-Database Language SQL", Jim Melton, Editor, International Organization for Standardization and American National Standards Institute, julio de 1992; Microsoft Corporation, "ODBC 2.0 Programmer's Reference and SDK Guide. Microsoft Open Database Standard for Microsoft Windows.TM y Windows NTTM, Microsoft Open Database Connectivity.TM. Software Development Kit," 1992, 1993, 1994 Microsoft Press, pp. 3- 30 y 41-56; ISO Working Draft, "SQL-Database Language-Part" 2: Foundation (SQL/Foundation), "CD9075-2: 199.chi.SQL, 11 de septiembre de 1997, y similares). En el documento WO 00/42559 titulado "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLU- TIONARY SIMULATIONS", de Selifonov y Stemmer, se encuentran detalles relevantes adicionales sobre las aplicaciones basadas en la web.

[0257] En algunas realizaciones, los métodos para explorar, seleccionar y/o desarrollar secuencias polinucleotídicas o polipeptídicas 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 de computación 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 involucra una nube se conoce como un entorno de computación en la nube. En algunas realizaciones, uno o más usuarios pueden acceder a las computadoras de la nube distribuidas a través de intranet y/o Internet. En algunas realizaciones, un usuario puede acceder remotamente, a través de un cliente web, computadoras de servidor que implementan los métodos para seleccionar y/o desarrollar variantes de proteínas descritas anteriormente.

[0258] 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 devolver al usuario. Una máquina virtual (VM) es una emulación basada en software de una computadora. Las máquinas virtuales pueden basarse en las especificaciones de una computadora hipotética o emular la arquitectura de la computadora y las funciones de una computadora del mundo real. La estructura y funciones de las máquinas virtuales son bien conocidas en la técnica. Normalmente, una VM está instalada en una plataforma que incluye hardware del sistema, y la VM en sí misma incluye hardware de sistema virtual y software invitado.

[0259] El hardware del sistema 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 de invitado. En algunas implementaciones, el software del sistema invitado incluye un sistema operativo invitado con controladores para dispositivos virtuales. En algunas implementaciones, las aplicaciones de invitados de la VM incluyen al menos una instancia de un sistema de detección de proteínas virtual como se describió anteriormente.

[0260] 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, incluyendo la máquina virtual un sistema de exploración virtual. En algunas realizaciones, el entorno de computación en la nube puede aprovisionar una VM en base a la solicitud del usuario. En algunas realizaciones, puede existir una VM en una imagen VM previamente almacenada, que puede almacenarse en un depósito de imágenes. El entorno de computación en la nube puede buscar y transferir la imagen a un servidor o a un sistema de usuario. El entorno de computación en la nube puede entonces iniciar la imagen en el servidor o sistema del usuario.

IX. EJEMPLOS

Ejemplo 1

[0261] El siguiente ejemplo ilustra un proceso de selección virtual de variantes de enzimas y enzimas en desarrollo de actividad catalítica deseada y selectividad implementando diversas realizaciones.

[0262] En resumen, el proceso implicó la creación de modelos de homología en 3 dimensiones de un panel real de enzimas y cribando prácticamente los miembros del panel de enzima para seleccionar una primera variante que (a) se acopló con el sustrato en una posición activa, (b) atracado en una conformación pro-S, y (c) tenía la energía de enlace total más baja (o puntaje de atraque) entre los que atracaron en posiciones activas y en una conformación pro-S. Luego, el proceso utilizó la primera variante como una cadena principal de la ronda 1, o secuencia parental, para crear una biblioteca de variante virtual de la ronda 1 usando técnicas de mutagénesis virtual para la evolución dirigida virtual. Luego, el proceso creó modelos de miembros de la biblioteca de variantes virtuales de la ronda 1, seleccionó la biblioteca de variante virtual de la ronda 1 y seleccionó una segunda variante como la red troncal de la ronda 2 utilizando métodos de selección similares a la selección de la red troncal de la ronda 1. El proceso también seleccionó variantes adicionales de la biblioteca de variantes virtuales de la ronda 1. Las variantes adicionales (a) acopladas con el sustrato en posiciones activas, y (b) tenían baja energía de enlace total (o puntuación de atraque)

entre las que atracan en posiciones activas. El proceso luego recombinó la red troncal de ronda 2 con las variantes adicionales para introducir la diversidad en una biblioteca de variantes de ronda 2. Finalmente, el proceso modeló computacionalmente, y seleccionó variantes, produciendo variantes de enzimas virtuales con actividad y selectividad mejoradas en comparación con las cadenas principales de ronda 1 y ronda 2.

[0263] Más específicamente, el proceso de ejemplo se inició mediante la creación de 194 modelos de homología de un panel real de enzimas. Estas enzimas catalizan un sustrato nativo que está relacionado estructural o funcionalmente con un sustrato deseado. El proceso conectó el sustrato deseado a los modelos de homología, y virtualmente seleccionó miembros del panel de enzima real para encontrar solo una variante que (a) se acopló con el sustrato deseado en una posición activa, y (b) atracó en una conformación pro-S. El enlace exitoso en una posición activa sugirió que era probable que el ligando experimentara una transformación catalítica o realizara alguna función deseada tal como la unión covalente con el sitio de unión. El acoplamiento del sustrato deseado y los miembros del panel se realizó mediante métodos de acoplamiento descritos en los detalles anteriores. Los restos funcionalmente relevantes del sustrato deseado se compararon con el sustrato nativo colocando los dos sustratos en las mismas coordenadas X, Y, Z en un espacio de acoplamiento. Si una posición del sustrato deseado estaba activa, pro-S o pro-R, se determinó por la distancia entre los restos del sustrato deseado y el sustrato nativo. El criterio de distancia se estableció en 1,25 Å para este ejemplo. El valor del criterio y las reglas (que requieren que la media, el mínimo, el máximo, etc. de las distancias sean menores que el criterio) pueden ajustarse en diferentes aplicaciones y en varias rondas de evolución dirigida.

[0264] Se encontró que esta variante se podría enlazar el sustrato en ambas conformaciones pro-R y pro-S. Se sospechaba que la variante podría no ser muy selectiva. Para derivar una enzima activa y selectiva S para el sustrato deseado, esta variante se seleccionó como una cadena principal ronda 1 para crear una biblioteca de variantes de ronda 1 mediante mutagénesis en la primera ronda de evolución dirigida *in silico.* Hubo 15 posiciones del sitio activo identificadas en esta cadena principal de la ronda 1, y 19 aminoácidos posibles para cada posición que serían diferentes de la variante de la cadena principal de la ronda 1, lo que equivale a 285 diferentes mutaciones puntuales posibles. En la evolución de la ronda 1, se generaron 1.000 mutantes para la biblioteca de variantes de la ronda 1, teniendo cada mutante un número aleatorio de mutaciones, seleccionándose el número aleatorio de una distribución gaussiana de la media = 4 y SD = 2. Las mutaciones fueron elegidas al azar de las 285 posibles mutaciones puntuales.

[0265] Luego, el proceso usó métodos de atraque y detección similares a los descritos anteriormente para el panel de enzimas real, con la excepción de que el criterio para determinar la actividad y la selectividad de las posiciones se estableció en un valor más estricto de 1 Å a diferencia de 1,25 Å. El proceso identificó una variante como la que comprende la mutación que tiene la energía de enlace total más baja entre todos los mutantes que se unirían en posiciones activas y pro-S. De hecho, la mutación en esta variante evitó que el sustrato se uniera en una conformación pro-R no deseada, que representa una mutación beneficiosa para la selectividad. El proceso, por lo tanto, seleccionó esta variante como la columna vertebral para una evolución dirigida a la ronda 2.

[0266] Sin embargo, la energía de unión de la cadena principal de la ronda 2 a 0,38303 kcal/mol era relativamente alta incluso en comparación con la determinada para la cadena principal de la ronda 1 (-4,005 kcal/mol), sugiriendo que la evolución podría mejorar las propiedades beneficiosas de la enzima. Una evolución dirigida a la ronda 2 se llevó a cabo *in silico* introduciendo 29 mutaciones en la red troncal de la ronda 2. Las 29 mutaciones se derivaron de 29 variantes de la biblioteca de la ronda 1 que tienen la energía de unión más baja entre todas las variantes obtenidas a partir de la evolución de la ronda 1. En la evolución de la ronda 2, se generaron 1.000 mutantes para producir la biblioteca de variantes de la ronda 2, teniendo cada mutante un número aleatorio de mutaciones, seleccionándose el número aleatorio de una distribución gaussiana de la media = 6 y SD = 4. Las mutaciones fueron elegidas al azar de las 29 posibles mutaciones derivadas de 29 variantes.

[0267] Entonces, el proceso utilizó métodos de atraque y detección similares a los descritos anteriormente para determinar que la mayoría de las variantes favorecían la unión del sustrato en una conformación pro-S deseada solamente, y al menos 10 variantes tenían una mejor energía de unión que las columnas de la ronda 1 y de la ronda 2. Véanse la Tabla 1 para las energías de enlace de las variantes mejoradas de la evolución de la ronda 2 y las columnas de la ronda 1 y ronda 2. Además de mostrar los datos de la Tabla 1, la Figura 5 muestra la selectividad de las 10 variantes mejoradas de la evolución de la ronda 2, así como las cadenas principales de la ronda 1 y de la ronda 2. La Figura ilustra que el cribado virtual del panel de enzima identificó primero la cadena principal de la ronda 1 que tenía una baja energía de unión, pero no era selectiva para S. El proceso luego mejoró la selectividad S usando la evolución dirigida *in silico* (mutagénesis), para obtener la cadena principal de la ronda 2. El proceso finalmente mejoró la unión del sustrato en la evolución de la ronda 2 a través de la recombinación, produciendo variantes de la enzima que tenían una alta afinidad con el sustrato deseado y que eran enantioselectivas.

Tabla 1. Energías de enlace de variantes de la evolución de la ronda 2	
Variantes	**Energía de enlace (kcal/mol)**
Rd2 Variante 10	-11,9
Rd2 Variante 9	-11,7
Rd2 Variante 8	-9,2
Rd2 Variante 7	-9,0
Rd2 Variante 6	-7,3
Rd2 Variante 5	-6,4
Rd2 Variante 4	-6,0
Rd2 Variante 3	-5,7
Rd2 Variante 2	-5,3
Rd2 Variante 1	-5,2
Rd2BB	0,4
Rd1BB	-4,0

[0268] La diversidad proporcionada en las dos rondas de evolución se generó mediante mutagénesis y recombinación, inspirada en operaciones genéticas biológicas. En algunas aplicaciones, el método de detección de proteína virtual se puede combinar con modelos de actividad de secuencia que guían los métodos de evolución dirigida. Se construyó un modelo de actividad de secuencia con técnicas de regresión lineal múltiple de acuerdo con los métodos descritos en la Patente de Estados Unidos Nº 7.783.428. En la Figura 6A, la energía de unión predicha del modelo de actividad de secuencia se traza frente a la energía observada obtenida por el sistema de exploración virtual para un conjunto de prueba de secuencias. La validación cruzada del modelo de actividad de secuencia se realizó al probar un conjunto de validación de secuencias omitidas del conjunto de prueba. El modelo representa el 90,9% de la varianza en el conjunto de prueba (R^2 = 0,909). Los datos de validación cruzada en la Figura 6B muestran que el modelo de actividad de secuencia fue preciso para predecir la energía de unión de las secuencias de mutaciones particulares en posiciones particulares, representando el 82,9% de la varianza en el conjunto de validación (R^2 = 0,829).

[0269] El modelo se puede usar para identificar aminoácidos para la mutagénesis. Entre otras formas de utilizar un modelo de actividad de secuencia para guiar la evolución dirigida, una forma se basa en los coeficientes de regresión para una mutación particular de un residuo específico en una posición específica, que refleja la contribución de la mutación a la actividad de la proteína. Específicamente, un proceso de evolución dirigida podría seleccionar las posiciones para la mutación evaluando los coeficientes de los términos del modelo de actividad de secuencia para identificar uno o más de los aminoácidos que contribuyen a la energía de unión sustancial calculada por el sistema de exploración virtual. Por ejemplo, en este ejemplo, la mutación 1 tiene un gran coeficiente positivo, lo que indica que la mutación 1 aumenta la actividad en gran medida. Véanse la Figura 6C. Por el contrario, la mutación 27 tiene un gran coeficiente negativo, lo que sugiere que esta mutación debe evitarse para obtener una actividad alta medida en la Figura 6C.

Ejemplo 2

[0270] El ejemplo 2 proporciona una validación experimental de filtrar virtualmente variantes de cetoreductasa para el enantiómero R de un alcohol quiral de una cetona pro-quiral, como la reacción mostrada en la parte superior de la Figura 7.

[0271] El proceso involucró la creación de modelos de homología tridimensional de dos paneles existentes de variantes de la enzima de cetoreductasa (formato de 96 pocillos para cada panel) y la detección virtual de los 192 miembros de los paneles de cetoreductasa para seleccionar variantes que (a) atracaran con el sustrato en una posición activa, (b) atracado en una conformación pro-R, y (c) tuvo puntaje de atraque favorable.

[0272] El proceso identificó 24 variantes que pueden conducir a posiciones activas y enérgicamente favorables, que pueden ser priorizadas para un mayor desarrollo y cribado. Para validar la utilidad y validez de los resultados de cribado *in silico* virtual, el proceso también realizó un cribado *in vitro* para los 192 miembros con un protocolo estándar, y los sustratos/productos se detectaron con cromatografía líquida de alta resolución (HPLC).

[0273] Los resultados se muestran en la Figura 7, donde el eje x es % de conversión calculado como ($\text{ÁreaPico}_{(R)\text{-alcohol}}$ + $\text{ÁreaPico}_{(S)\text{-alcohol}}$) ÷ ($\text{ÁreaPico}_{(R)\text{-alcohol}}$ + $\text{ÁreaPico}_{(S)\text{-alcohol}}$ + $\text{ÁreaPico}_{\text{cetona}}$) X 100% y el eje y es % e.e. hacia el producto R deseado (un índice de enantioselectividad) calculado como ($\text{ÁreaPico}_{(R)\text{-alcohol}}$ - $\text{ÁreaPico}_{(S)\text{-alcohol}}$) ÷ ($\text{ÁreaPico}_{(R)\text{-alcohol}}$ + $\text{ÁreaPico}_{(S)\text{-alcohol}}$) X 100%. Las 24 variantes priorizadas por el cribado virtual se destacaron como Red Square y las variantes restantes se destacaron como Blue Diamond. Los resultados sugieren: 1) el cribado virtual puede ayudar a determinar si una conversión deseada es factible con un conjunto de variantes de enzimas antes de cualquier cribado *in vitro;* 2) una buena cantidad de variantes predichas de hecho dieron una alta actividad

(% de conversión) y enantioselectividad (% e.e.), a pesar del hecho de que un sustrato tan pequeño y flexible se considera habitualmente como un desafío para el modelado. Por lo tanto, el cribado virtual puede filtrar reacciones muy poco probables para el cribado *in vitro* y seleccionar menos muestras para análisis (24 frente a 192 en este caso), lo que puede llevar a ahorros significativos en tiempo y costes.

Ejemplo 3

[0274] Ejemplo 3 proporciona una validación experimental de evolución dirigida virtual de transaminasa para reducción C=O estereoselectiva a CH-NH_2, como la reacción que se muestra en la parte superior de la Figura 8.

[0275] El proceso implicó la creación de modelos en 3 dimensiones de homología de 228 secuencias virtuales de mutagénesis saturada *in silico* de 12 posiciones del sitio activo de la columna (12 posiciones X 19 AA/posición = 228 variantes, 1 mutación/variante) y virtualmente seleccionando las 228 variantes virtuales para seleccionar variantes que (a) se acoplaron con el sustrato en una posición activa, (b) se acoplaron en una conformación que conduce a la estereoselectividad deseada, y (c) tenía la energía de unión total más baja entre los que se acoplaron en posiciones activas y en una conformación específica.

[0276] El proceso identificó 12 variantes o 12 mutaciones que pueden conducir a posiciones activas y energéticamente favorables. Las 12 mutaciones se usaron para sintetizar una biblioteca, que se exploró *in vitro.* El cribado *in vitro* se llevó a cabo para 360 variantes (una o más mutaciones por variante) con un protocolo patentado. El sustrato/productos se detectaron con HPLC.

[0277] Los resultados para las mejores variantes de cribado *in vitro* se muestran en la Figura 8, donde el eje x es las muestras filtradas, y el eje y es FIOPC definido como mejora de pliegue sobre control positivo y se calculó como (%$Conversión_{Variante}$ - %$Conversión_{ControlNegativo}$) ÷ (%$Conversión_{ControlPositivo}$ - %$Conversión_{ControlNegativo}$) X 100%. El Control Positivo es la columna del cribado virtual y el cribado *in vitro,* y el Control Negativo es el vector vacío sin enzima.

[0278] La filtración de la biblioteca *in vitro* dio como resultado que el 13% de las variantes tenía un FIOPC >1,5 y 5,3% con un FIOPC >2. El resultado más alto tenía un FIOPC de 2,4. El cribado virtual puede, por lo tanto, filtrar las mutaciones perjudiciales para el cribado *in vitro* y ayudar a diseñar bibliotecas más específicas, lo que puede llevar a ahorros significativos en tiempo y costes. Por ejemplo, si tuviéramos que hacer la etapa de mutagénesis saturada *in vitro,* al menos se tendrá que explorar otras 800 variantes.

REIVINDICACIONES

1. Un método, implementado usando un sistema informático que incluye uno o más procesadores y memoria del sistema, para seleccionar una pluralidad de diferentes variantes de enzima para actividad con un sustrato, donde la pluralidad de diferentes variantes de enzima comprende al menos diez variantes diferentes, y las variantes de enzima comprenden sitios activos que difieren de otro por al menos una mutación en la secuencia de aminoácidos del sitio activo, comprendiendo el método:

(a) crear o recibir un modelo estructural para cada una de la pluralidad de diferentes variantes de enzima, en donde cada modelo estructural contiene una representación computacional tridimensional de un sitio activo de una variante de enzima;
(b) para cada variante enzimática, acoplamiento, por el sistema informático, una representación computacional del sustrato a la representación computacional tridimensional del sitio activo de la variante enzimática, en donde el acoplamiento (i) genera una pluralidad de posiciones del sustrato en el sitio activo, en donde una posición comprende una posición u orientación del sustrato con respecto al sitio activo de la variante enzimática, e (ii) identifica posiciones energéticamente favorables del sustrato en el sitio activo, en donde una posición energéticamente favorable es una posición que tiene una energía que sea favorable para la unión entre el sustrato y la variante enzimática;
(c) para cada posición energéticamente favorable, determinar si la posición está activa, en donde una posición activa cumple una o más restricciones para que el sustrato experimente una reacción catalítica en el sitio activo; y
(d) seleccionar al menos una de las variantes de enzima que tiene un sitio activo en el que el sustrato tiene una o más posiciones activas como se determina en (c).

2. El método de la reivindicación 1, que comprende además:

(i) seleccionar al menos una variante de enzima seleccionada en (d) contra el sustrato produciendo una reacción química; o
(ii) sintetizar al menos una variante de enzima seleccionada en (d).

3. El método de cualquiera de las reivindicaciones precedentes, en el que la representación computacional del sustrato:

(i) representa una especie a lo largo de la coordenada de reacción para la actividad de la enzima, seleccionándose la especie del sustrato, una reacción intermedia del sustrato, o un estado de transición del sustrato; o
(ii) es un modelo tridimensional del sustrato.

4. El método de cualquiera de las reivindicaciones precedentes, en el que la pluralidad de variantes de enzima:

(i) comprende un panel de enzimas que puede convertir múltiples sustratos y en donde los miembros del panel poseen al menos una mutación con respecto a una secuencia de referencia, y opcionalmente donde al menos una mutación es una mutación de único residuo en el sitio activo de la enzima; o
(ii) comprende una o más enzimas que pueden catalizar una reacción química seleccionada entre la oxidorreducción, la transfección, la hidrólisis, la isomerización, la ligación y la ruptura del enlace químico mediante una reacción distinta a la hidrólisis, oxidación o reducción.

5. El método de la reivindicación 4 (ii), en el que:

(i) la enzima se selecciona de oxidorreductasa, transferasa, hidrolasa, isomerasa, ligasa y liasa; o
(ii) la pluralidad de variantes comprende una o más enzimas que pueden catalizar una reacción química seleccionada entre reducción de cetona, transaminación, oxidación, hidrólisis de nitrilo, reducción de imina, reducción de enona, hidrólisis de acilo y deshalogenación de halohidrina, y en donde opcionalmente se selecciona la enzima de reductasa de cetona, transaminasa, citocromo P450, monooxigenasa Baeyer-Villiger, monoaminooxidasa, nitrilasa, reductasa de imina, reductasa de enona, acilasa y deshalogenasa de halohidrina.

6. El método de cualquiera de las reivindicaciones precedentes, en el que la pluralidad de variantes comprende al menos aproximadamente cien variantes diferentes o al menos aproximadamente mil variantes diferentes.

7. El método de cualquiera de las reivindicaciones precedentes, en el que las representaciones computacionales de sitios activos se proporcionan a partir de modelos de homología tridimensional para la pluralidad de variantes, comprendiendo el método opcionalmente además la producción de dichos modelos de homología tridimensional para la pluralidad de variantes.

8. El método de cualquiera de las reivindicaciones precedentes, en el que el método se aplica para seleccionar una pluralidad de sustratos.

9. El método de cualquiera de las reivindicaciones precedentes, que comprende además la identificación de las restricciones para que el sustrato experimente la transformación química catalizada identificando una o más posiciones de un sustrato nativo, un intermedio de reacción del sustrato nativo, o un estado de transición del sustrato nativo cuando el sustrato nativo se somete a la transformación química catalizada por una enzima de tipo salvaje.

10. El método de cualquiera de las reivindicaciones precedentes, en el que:

(i) las restricciones comprenden uno o más de los siguientes: restricciones de posición, distancia, ángulo y torsión;
(ii) las restricciones comprenden una distancia entre un resto particular en el sustrato y un resto particular o resto en el sitio activo;
(iii) las restricciones comprenden una distancia entre un resto particular en el sustrato y un resto particular o resto en un cofactor; y/o
(iv) las restricciones comprenden una distancia entre un resto particular en el sustrato y un sustrato nativo posicionado idealmente en el sitio activo.

11. El método de cualquiera de las reivindicaciones precedentes, comprendiendo el método además la aplicación de un conjunto de una o más restricciones enzimáticas a la pluralidad de variantes enzimáticas, donde una o más restricciones enzimáticas son similares a las restricciones de una enzima natural cuando un sustrato nativo se somete a una transformación química catalizada en presencia de la enzima de tipo salvaje.

12. El método de cualquiera de las reivindicaciones precedentes, en el que:

(i) la pluralidad de posiciones del sustrato se obtiene mediante una o más operaciones de acoplamiento seleccionadas del grupo que consiste en: dinámica molecular de alta temperatura, rotación aleatoria, refinamiento por anillado simulado basado en grillas, minimización de campos de fuerza completa o de grilla, y cualquier combinación de los mismos;
(ii) la pluralidad de posiciones del ligando comprende al menos aproximadamente 10 posiciones, al menos aproximadamente 20 posiciones, al menos aproximadamente 50 posiciones, o al menos aproximadamente 100 posiciones, del sustrato en el sitio activo; o
(iii) al menos una variante enzimática tiene actividad catalítica y/o selectividad deseadas.

13. El método de cualquiera de las reivindicaciones precedentes, en el que la selección en (d) comprende:

(i) identificar variantes que se determina que tienen un gran número de posiciones activas en comparación con otras variantes; o
(ii) clasificar las variantes por una o más de las siguientes: el número de posiciones activas que tienen las variantes, puntajes de atraque de las posiciones activas, donde opcionalmente los puntajes de atraque se basan en la fuerza de van de Waals y la interacción electrostática, y energías de enlace de las posiciones activas, en donde opcionalmente las energías de enlace se basan en una o más de las siguientes: fuerza de van der Waals, interacción electrostática y energía de solvatación; y

seleccionar variantes basadas en sus rangos.

14. Un producto de programa de computadora que comprende uno o más medios de almacenamiento no transitorios legibles por computadora, estando instrucciones ejecutables por computadora almacenadas en ellos, ejecutadas por uno o más procesadores de un sistema informático, dan lugar a que el sistema informático aplique un método como se ha indicado en las reivindicaciones 1-13.

15. Un sistema que comprende:

uno o más procesadores;
memoria del sistema; y
en donde uno o más procesadores y memoria están configurados para implementar un método según se enumera en cualquiera de las reivindicaciones 1-13.

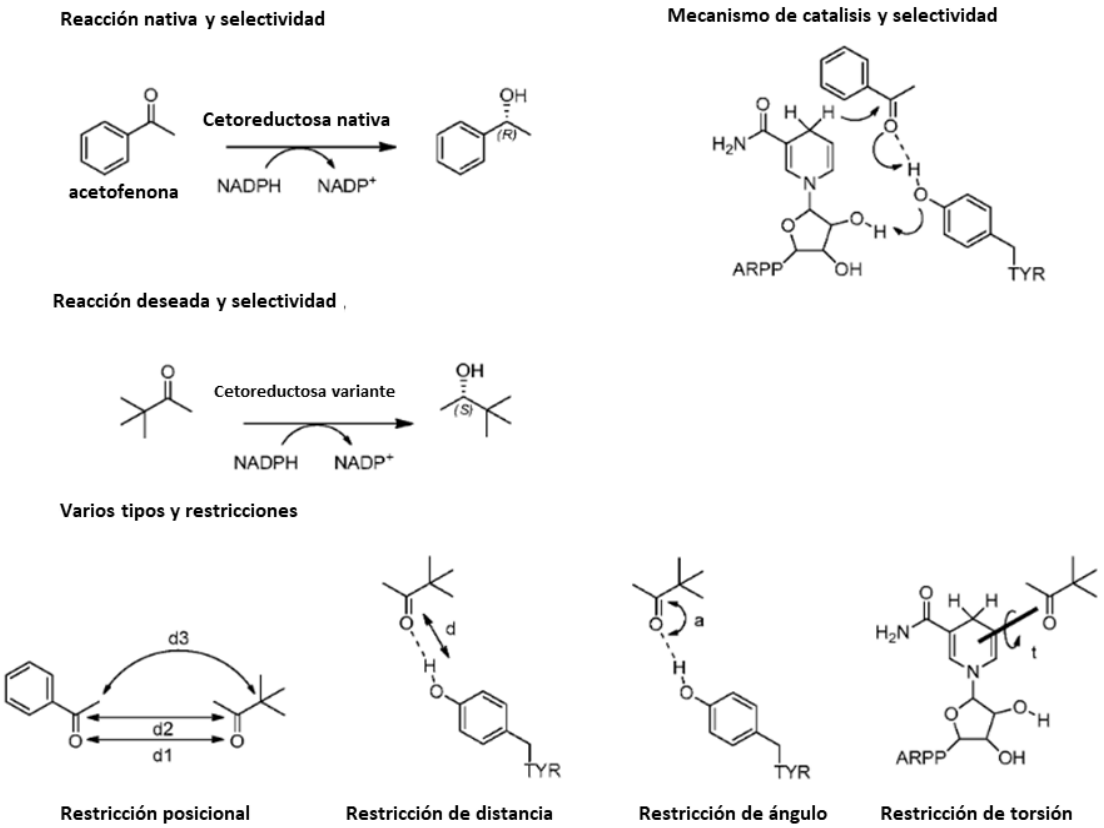
Reacción nativa y selectividad
acetofenona
Cetoreductosa nativa
NADPH
NADP+
OH
(R)
Mecanismo de catalisis y selectividad
H2N
H
H
O
H
N
O
H
O
H
ARPP
OH
TYR
Reacción deseada y selectividad
Cetoreductosa variante
NADPH
NADP+
OH
(S)
Varios tipos y restricciones
d3
d2
d1
O
O
d
a
t
Restricción posicional
Restricción de distancia
Restricción de ángulo
Restricción de torsión

FIG. 1

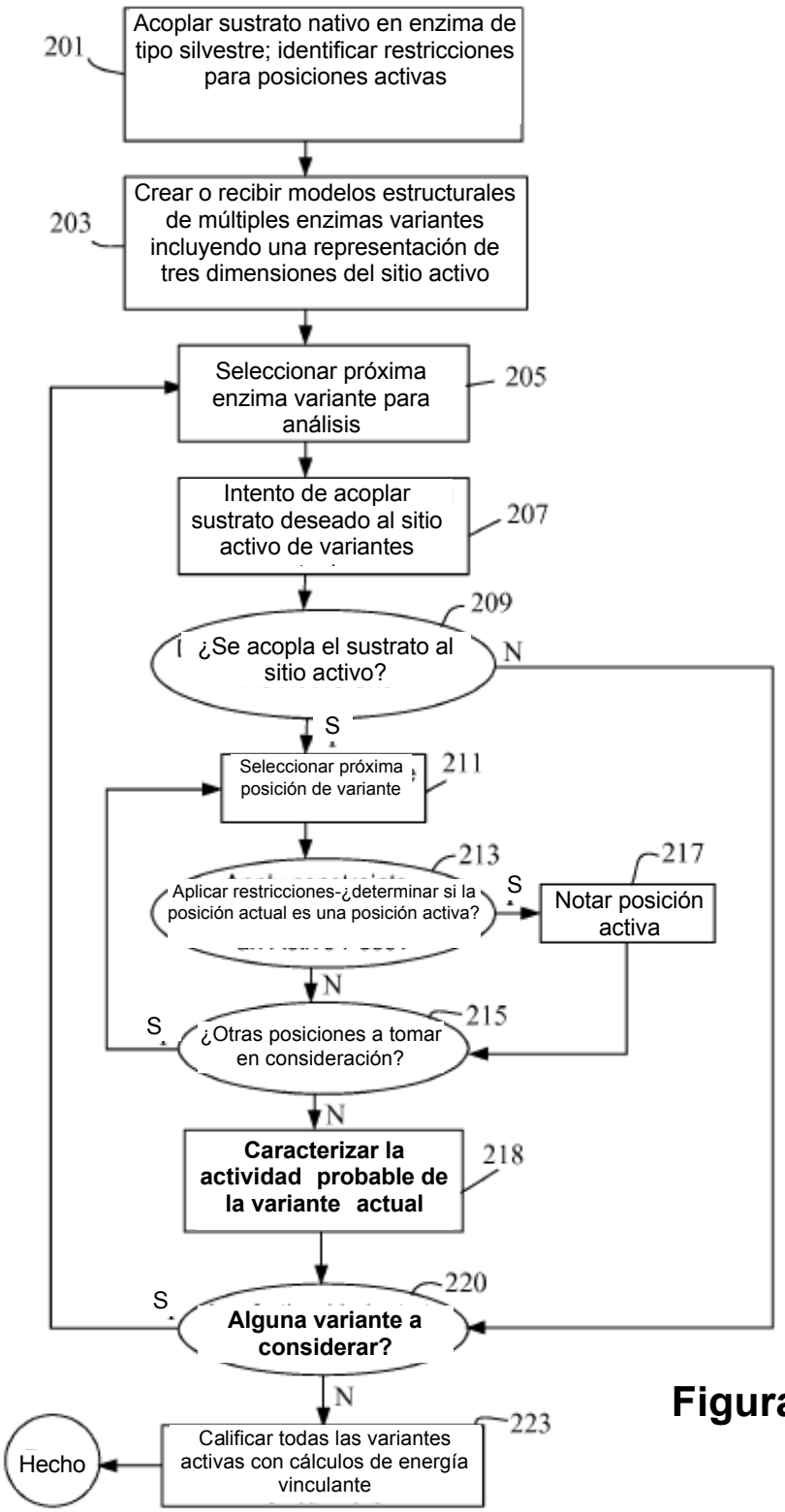
Acoplar sustrato nativo en enzima de tipo silvestre; identificar restricciones para posiciones activas
201
Crear o recibir modelos estructurales de múltiples enzimas variantes incluyendo una representación de tres dimensiones del sitio activo
203
Seleccionar próxima enzima variante para análisis
205
Intento de acoplar sustrato deseado al sitio activo de variantes
207
¿Se acopla el sustrato al sitio activo?
209
N
S
Seleccionar próxima posición de variante
211
Aplicar restricciones-¿determinar si la posición actual es una posición activa?
213
S
Notar posición activa
217
N
¿Otras posiciones a tomar en consideración?
215
S
N
Caracterizar la actividad probable de la variante actual
218
Alguna variante a considerar?
220
S
N
Calificar todas las variantes activas con cálculos de energía vinculante
223
Hecho

Figura 2

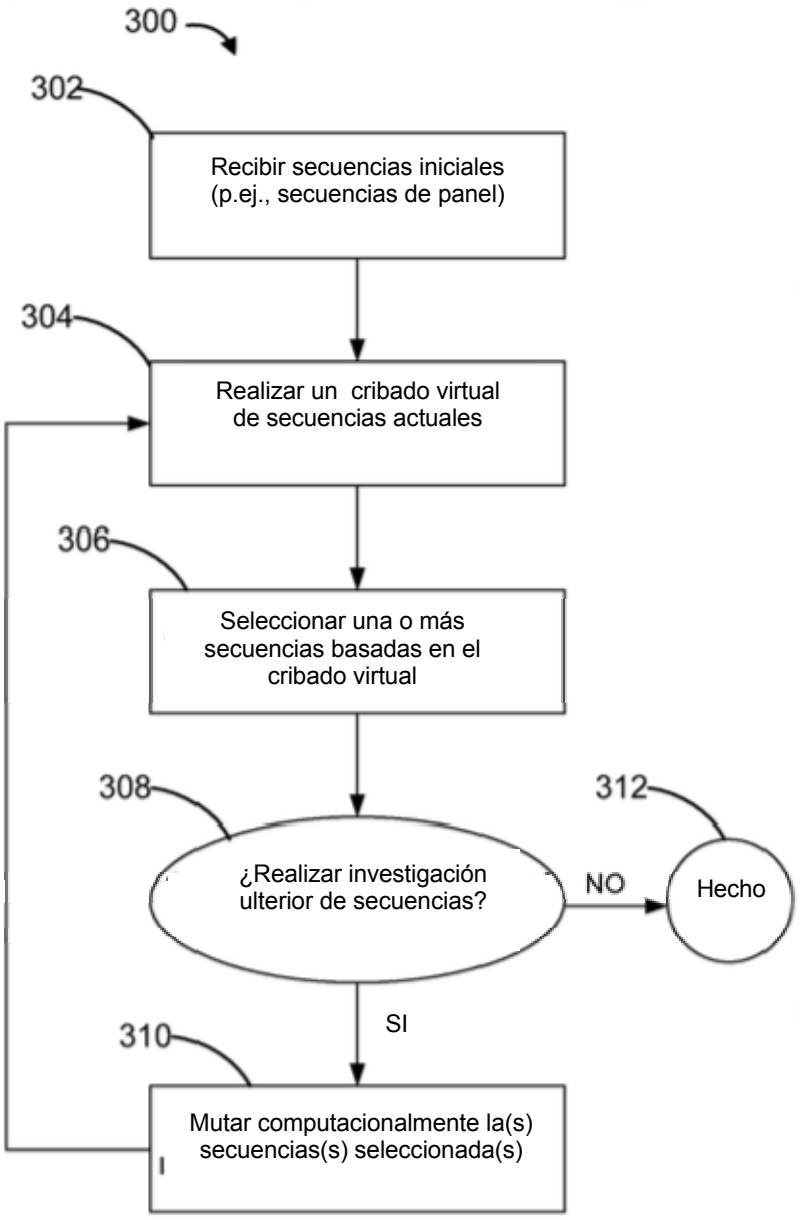
300
302
Recibir secuencias iniciales (p.ej., secuencias de panel)
304
Realizar un cribado virtual de secuencias actuales
306
Seleccionar una o más secuencias basadas en el cribado virtual
308
¿Realizar investigación ulterior de secuencias?
NO
312
Hecho
SI
310
Mutar computacionalmente la(s) secuencias(s) seleccionada(s)

FIG. 3A

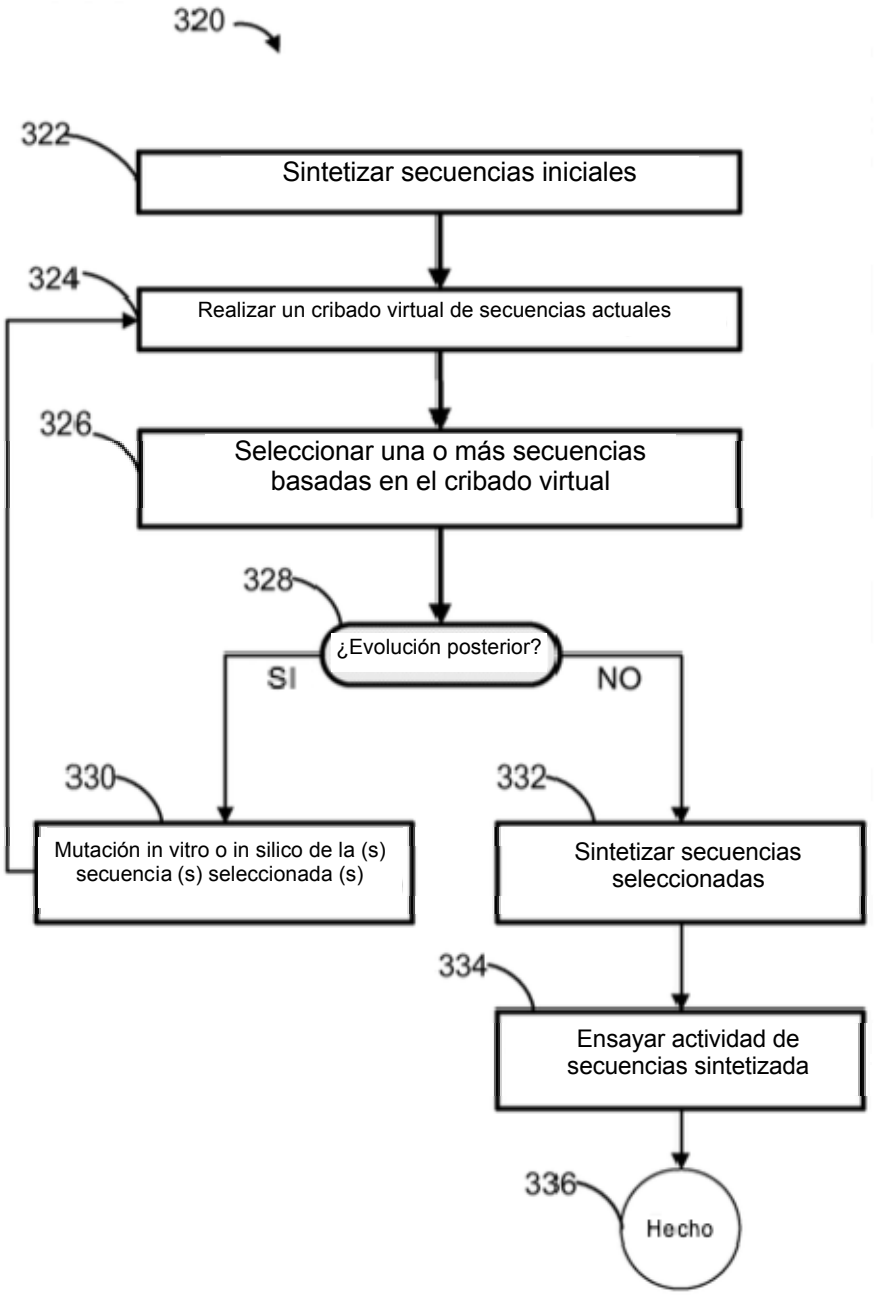
320
322
Sintetizar secuencias iniciales
324
Realizar un cribado virtual de secuencias actuales
326
Seleccionar una o más secuencias basadas en el cribado virtual
328
¿Evolución posterior?
SI
NO
330
Mutación in vitro o in silico de la (s) secuencia (s) seleccionada (s)
332
Sintetizar secuencias seleccionadas
334
Ensayar actividad de secuencias sintetizada
336
Hecho

FIG. 3B

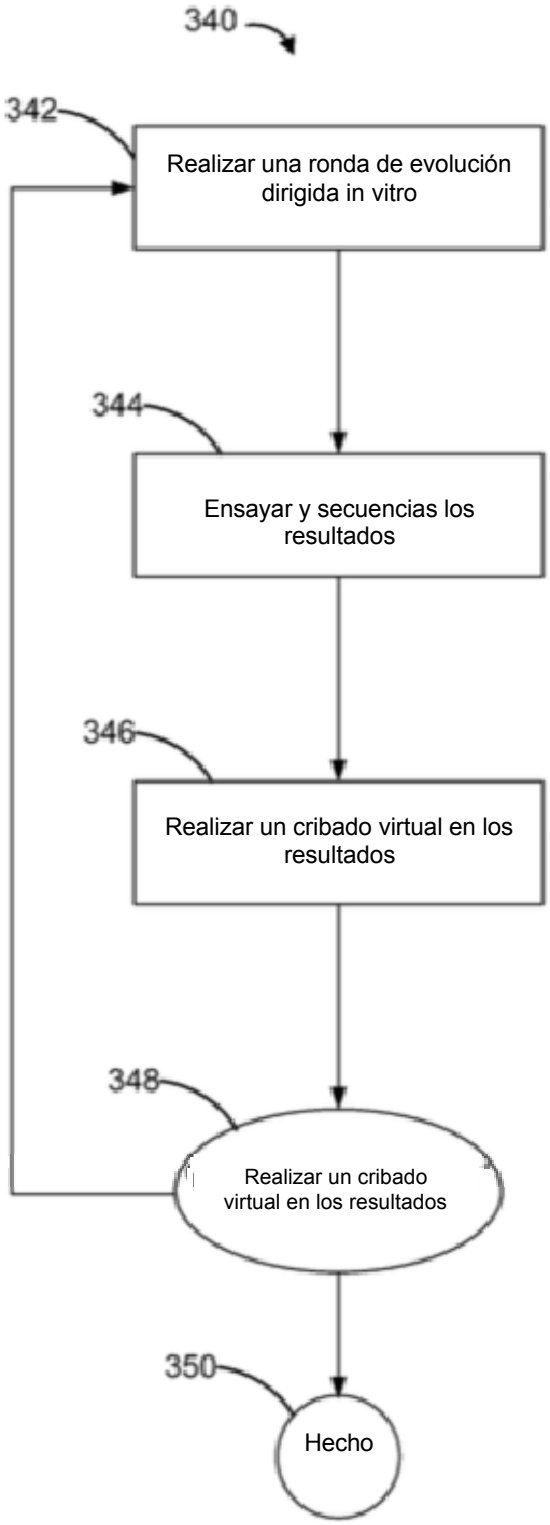
340
342
Realizar una ronda de evolución dirigida in vitro
344
Ensayar y secuencias los resultados
346
Realizar un cribado virtual en los resultados
348
Realizar un cribado virtual en los resultados
350
Hecho

FIG. 3C

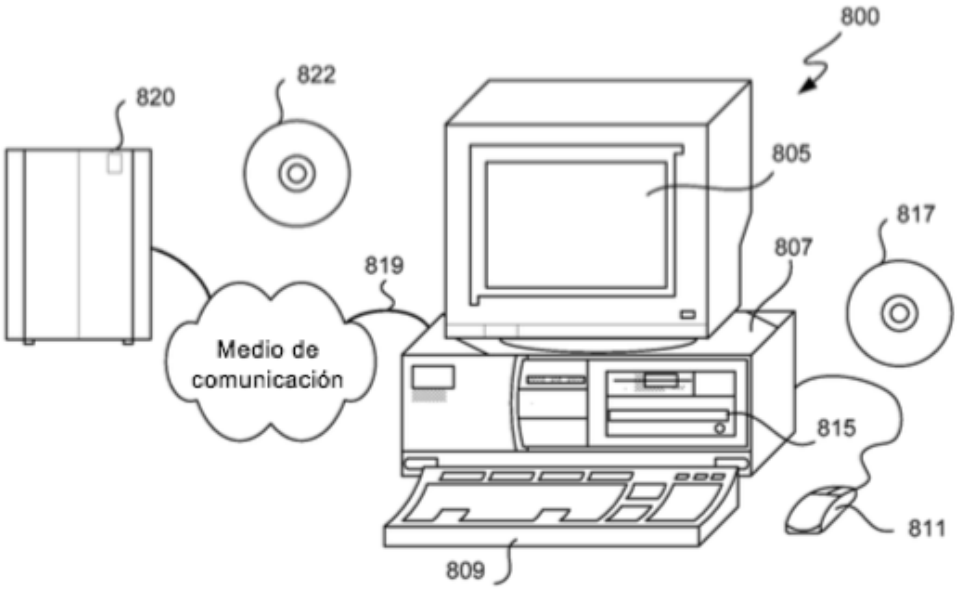

FIG. 4

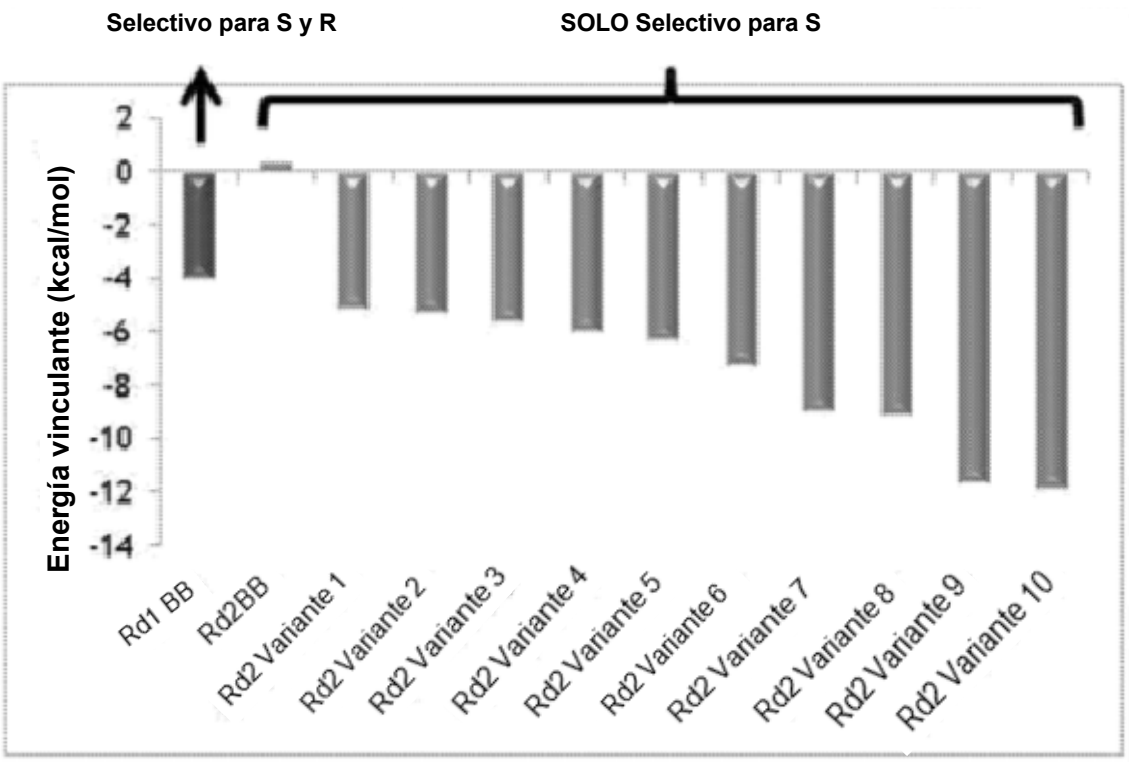
Selectivo para S y R
SOLO Selectivo para S
Energía vinculante (kcal/mol)
2
0
-2
-4
-6
-8
-10
-12
-14
Rd1 BB
Rd2BB
Rd2 Variante 1
Rd2 Variante 2
Rd2 Variante 3
Rd2 Variante 4
Rd2 Variante 5
Rd2 Variante 6
Rd2 Variante 7
Rd2 Variante 8
Rd2 Variante 9
Rd2 Variante 10

FIG. 5

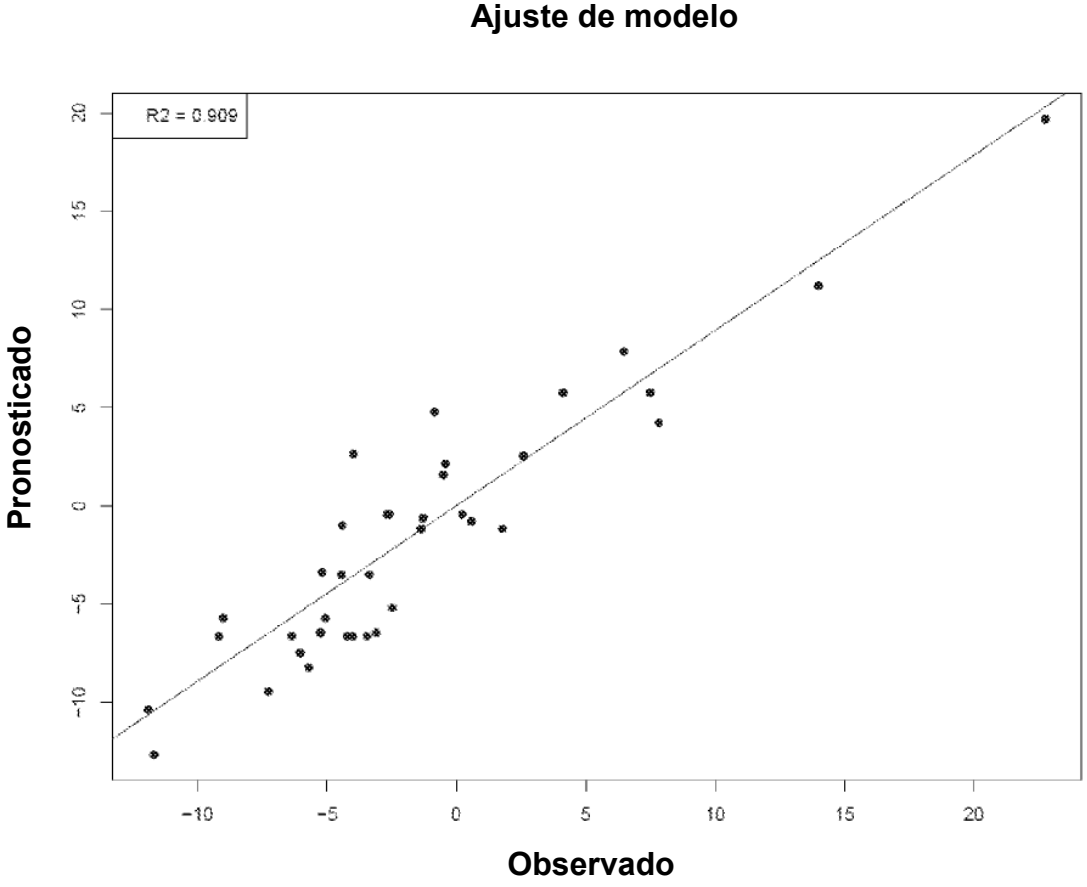
Ajuste de modelo
R2 = 0.909
Pronosticado
Observado
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20

FIG. 6A

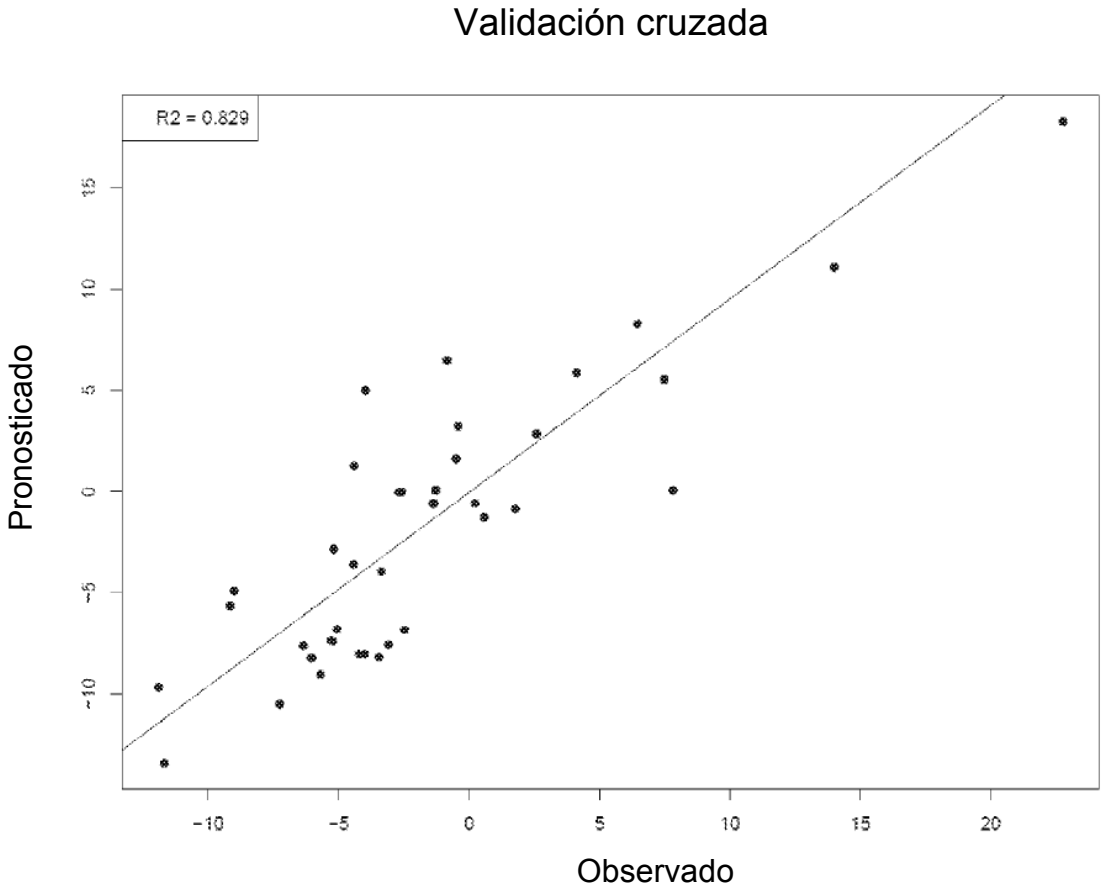
Validación cruzada
R2 = 0.829
Pronosticado
Observado
−10
−5
0
5
10
15
20

FIG. 6B

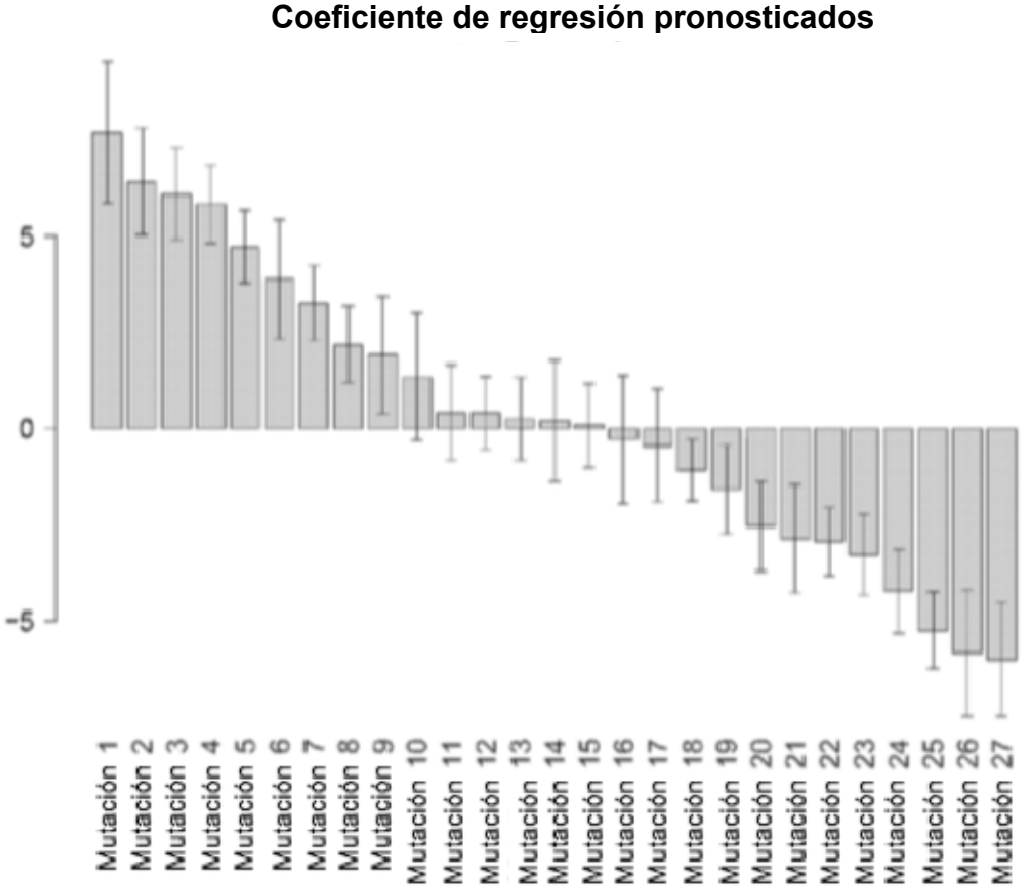
Coeficiente de regresión pronosticados
5
0
-5
Mutación 1
Mutación 2
Mutación 3
Mutación 4
Mutación 5
Mutación 6
Mutación 7
Mutación 8
Mutación 9
Mutación 10
Mutación 11
Mutación 12
Mutación 13
Mutación 14
Mutación 15
Mutación 16
Mutación 17
Mutación 18
Mutación 19
Mutación 20
Mutación 21
Mutación 22
Mutación 23
Mutación 24
Mutación 25
Mutación 26
Mutación 27

FIG. 6C

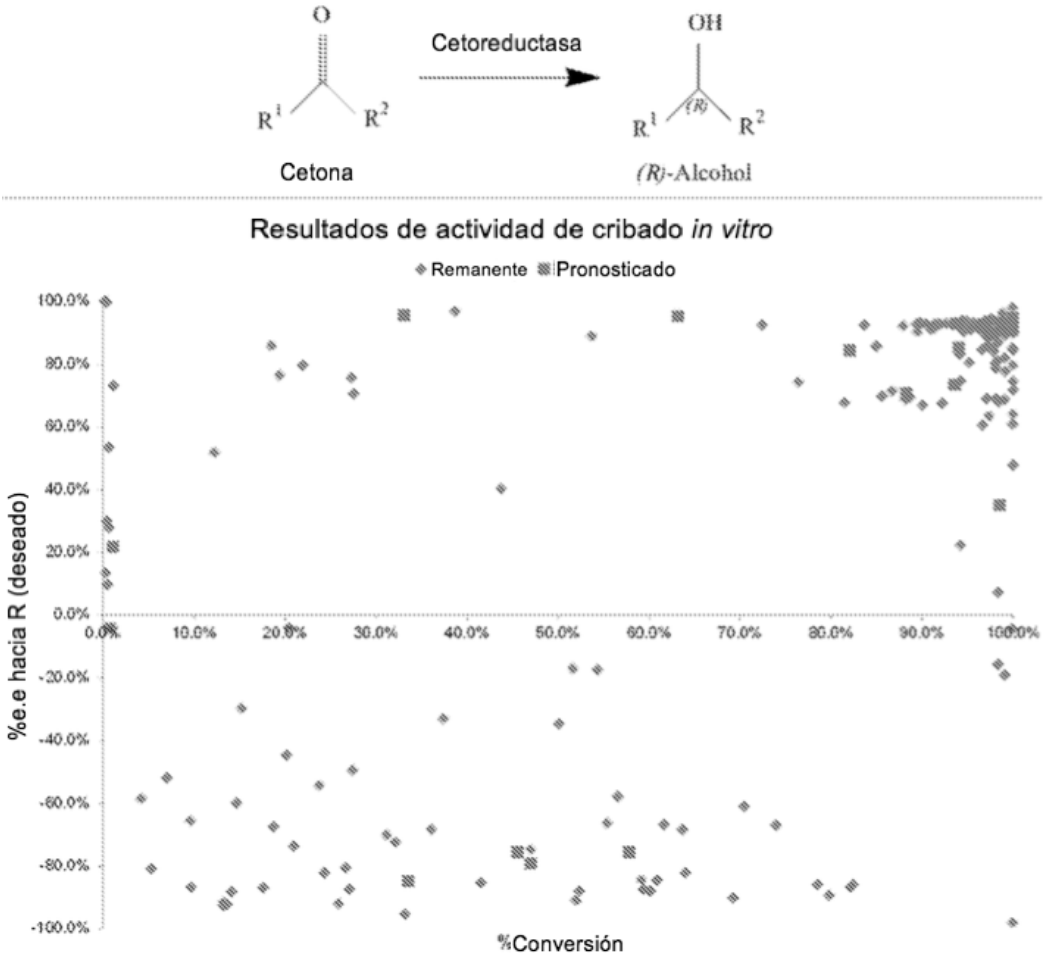
O
Cetoreductasa
OH
R1
R2
(R)
R1
R2
Cetona
(R)-Alcohol
Resultados de actividad de cribado in vitro
Remanente
Pronosticado
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
-60.0%
-80.0%
-100.0%
%e.e hacia R (deseado)
0.0%
10.0%
20.0%
30.0%
40.0%
50.0%
60.0%
70.0%
80.0%
90.0%
100.0%
%Conversión

FIG. 7

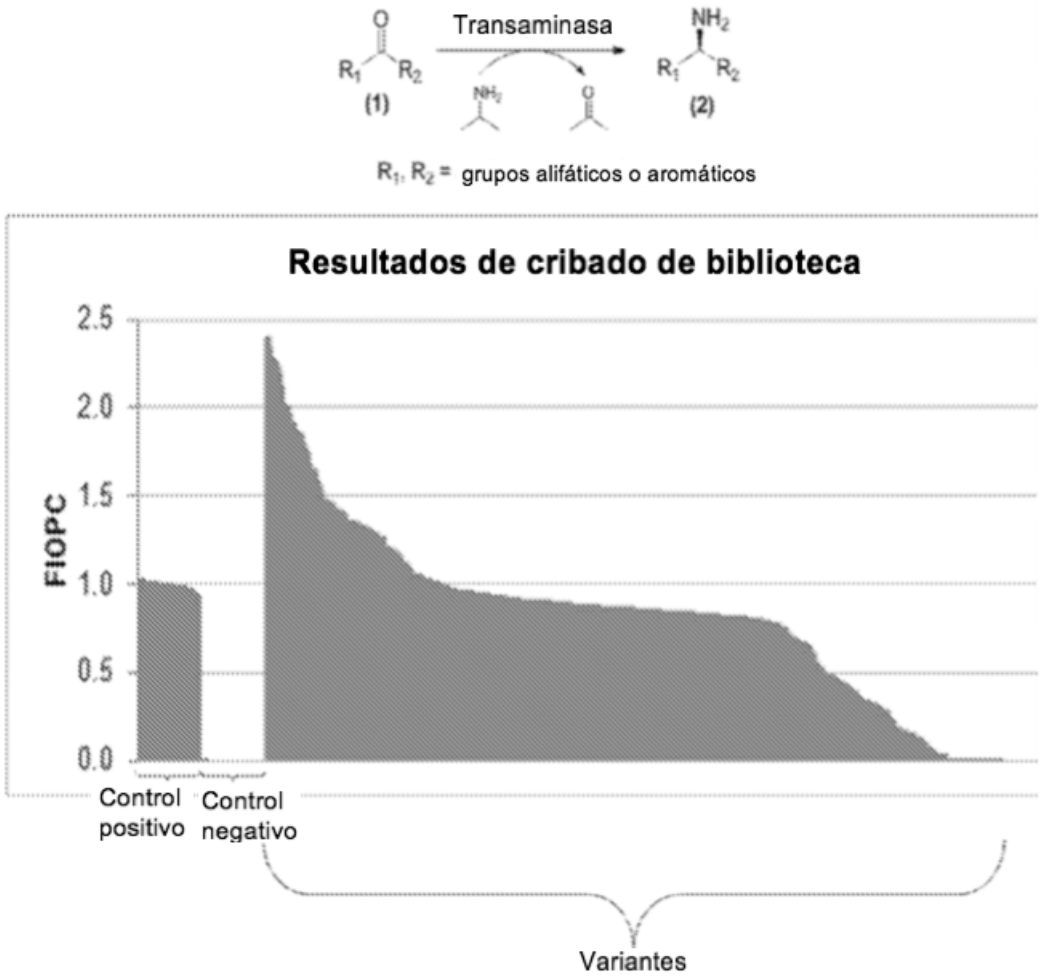
Transaminasa
(1)
(2)
R1, R2 = grupos alifáticos o aromáticos
Resultados de cribado de biblioteca
FIOPC
2.5
2.0
1.5
1.0
0.5
0.0
Control positivo
Control negativo
Variantes

FIG. 8