(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 740 323**

(51) Int. CI.:

G16B 5/20 (2009.01)
G16B 40/20 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **28.05.2014 PCT/US2014/039832**

(87) Fecha y número de publicación internacional: **04.12.2014 WO14193982**

(96) Fecha de presentación y número de la solicitud europea: **28.05.2014 E 14804788 (9)**

(97) Fecha y número de publicación de la concesión europea: **24.07.2019 EP 3005199**

(54) Título: **Redes de respuesta a paradigma de fármaco**

(30) Prioridad:

28.05.2013 US 201361828145 P
20.12.2013 US 201361919289 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
05.02.2020

(73) Titular/es:

FIVE3 GENOMICS, LLC (100.0%)
101 Cooper Street
Santa Cruz, California 95060, US

(72) Inventor/es:

BENZ, STEPHEN CHARLES y
SZETO, CHRISTOPHER

(74) Agente/Representante:

MARTÍN SANTOS, Victoria Sofia

DESCRIPCIÓN

Redes de respuesta a paradigma de fármaco.

Campo de la invención

El campo de la invención es el modelado computacional y el uso de patrones de modelos, especialmente en lo que se refiere a la modulación *in sílico* de patrones de modelos para identificar elementos de patrones útiles para el desarrollo de recomendaciones para tratamientos.

Antecedentes

La descripción de los antecedentes incluye información que puede ser útil para comprender la presente invención. No se admite que ninguna de la información proporcionada en este documento sea técnica anterior o relevante para la invención reivindicada actualmente, o que cualquier publicación a la que se haga referencia específica o implícita sea técnica anterior.

En la técnica se conocen diversos sistemas y métodos de modelado computacional de patrones. Por ejemplo, algunos algoritmos (por ejemplo, GSEA, SPIA y PathOlogist) son capaces de identificar con éxito patrones alterados de interés utilizando patrones seleccionados de la literatura. Incluso más, otras herramientas han construido gráficos causales a partir de interacciones seleccionadas en la literatura y han usado estos gráficos para explicar los perfiles de expresión. Algoritmos como ARACNE, MINDy y CONEXIC incorporan información transcripcional de genes (y número de copia, en el caso de CONEXIC) para identificar los posibles factores transcripcionales en un conjunto de muestras de cáncer. Sin embargo, estas herramientas no intentan agrupar diferentes impulsores en redes funcionales que identifiquen objetivos singulares de interés. Algunos algoritmos de patrones más nuevos, como NetBox y Mutual Exclusivity Modules in Cancer (MEMo) intentan resolver el problema de la integración de datos en cáncer para identificar redes a través de múltiples tipos de datos que son clave para el potencial oncogénico de las muestras.

Si bien dichas herramientas permiten al menos una integración limitada a través de los patrones para encontrar una red, generalmente no brindan información reguladora y la asociación de dicha información con uno o más efectos en los patrones o en la red de patrones relevantes. Del mismo modo, GIENA busca interacciones de genes desreguladas dentro de un solo patrón biológico, pero no tiene en cuenta la topología del patrón o el conocimiento previo sobre la dirección o la naturaleza de las interacciones. Además, debido a la naturaleza relativamente incompleta de estos sistemas de modelado, el análisis predictivo es a menudo imposible, especialmente cuando se investigan las interacciones de múltiples patrones y/o elementos del patrón.

Más recientemente, se han descrito varios sistemas y métodos mejorados para obtener *in sílico* patrones de modelos de patrones *in vivo*, y sistemas y métodos ejemplares se describen en el documento WO 2011/139345 y WO 2013/062505. El refinamiento adicional de tales modelos se proporcionó en el documento WO 2014/059036 (denominado colectivamente en el presente documento "PARADIGM") que divulga métodos para ayudar a identificar correlaciones cruzadas entre diferentes patrones de elementos y patrones. Si bien dichos modelos proporcionan información valiosa, por ejemplo, sobre las interconexiones de varios patrones de señalización y el flujo de señales a través de varios patrones, numerosos aspectos del uso de dicho modelado no han sido apreciados ni reconocidos. J. Wang et al. (Briefings in Bioinformatics., Vol. 13, no. 4, 27-ene-2012) desvela el uso de ensayos biológicos de alto rendimiento para descifrar patrones aberrantes y actividades de red. En particular, esta revisión proporciona ejemplos específicos en los que se han aplicado datos de alto rendimiento para identificar relaciones entre enfermedades y patrones aberrantes. *Michael P. Menden, et al. (PLOS One, vol. 28, No. 4, 1 de enero de 2013)* divulga modelos de aprendizaje automáticos para predecir la respuesta de las líneas celulares de cáncer al tratamiento farmacológico, cuantificadas a través de los valores de IC50, basándose tanto en las características genómicas de las líneas celulares como en las propiedades químicas de los medicamentos considerados.

Cuando una definición o uso de un término en una referencia es inconsistente o contrario a la definición de ese término provisto en la presente, la definición de ese término aquí provisto se aplica y la definición de ese término en la referencia no se aplica.

Por lo tanto, todavía existe la necesidad de proporcionar modelos y métodos computacionales mejorados para predecir *in sílico* la respuesta de uno o más patrones en una célula o tejido enfermo a una afección

simulada (por ejemplo, intervención terapéutica simulada) para ayudar a predecir un resultado terapéutico deseado.

Sumario de la invención

El presente ámbito de la invención está dirigido a dispositivos, sistemas y métodos para la predicción de un resultado terapéutico utilizando datos ómnicos obtenidos de una muestra de paciente y a priori patrones de modelos. En aspectos preferidos, la predicción de resultados terapéuticos se basa en la modulación *in sílico* de un modelo de patrón para simular un enfoque terapéutico, y el resultado de la simulación se emplea para preparar una recomendación de tratamiento.

En un aspecto del ámbito de la invención, se proporciona un método *in sílico* de acuerdo con la reivindicación 1. Cuando se desee o sea necesario, se contempla que los sistemas y métodos en el presente documento también incluyan una etapa adicional de preprocesamiento de los conjuntos de datos (por ejemplo, característica selección, transformación de datos, transformación de metadatos y/o división en conjuntos de datos de capacitación y validación).

Habitualmente, al menos uno de los conjuntos de datos distintos se genera a partir de una muestra de un paciente diagnosticado con una enfermedad neoplásica, a la vez que uno o más conjuntos de datos adicionales se generan a partir de cultivos celulares distintos que contienen células que no son del paciente. Cabe señalar que las células de los cultivos celulares son del mismo tipo neoplásico que la enfermedad neoplásica del paciente (por ejemplo, varias líneas celulares de cáncer de mama no derivadas del paciente y células o tejidos de cáncer de mama). Además, se debe apreciar que el paciente no habrá sido tratado por la enfermedad neoplásica. Visto desde otra perspectiva, los sistemas y métodos contemplados son adecuados para predecir combinaciones de medicamentos adecuadas para un resultado optimizado basado en datos ómicos del paciente incluso antes de que comience el tratamiento. Aunque no se limita al tema inventivo, generalmente se prefiere que se generen datos de salida que comprendan una recomendación de tratamiento para el paciente. Por lo tanto, los métodos contemplados también incluirán una etapa de identificación de un fármaco que se dirige al elemento del patrón determinante cuando el cambio de estado excede un umbral predeterminado.

Visto desde una perspectiva diferente, debe apreciarse que la pluralidad de células enfermas distintas diferirán entre sí con respecto a la sensibilidad de las células a un fármaco (u otra modalidad de tratamiento, incluyendo radiación, tratamiento térmico, etc.). Por ejemplo, un primer conjunto de células enfermas distintas puede ser sensible al tratamiento con un fármaco, mientras que un segundo conjunto de células enfermas distintas puede ser resistente al tratamiento con el fármaco.

Con respecto a los datos ómicos, todos los datos ómicos conocidos se consideran adecuados y los datos ómicos preferidos incluyen especialmente datos de número de copia de genes, datos de mutación de genes, datos de metilación de genes, datos de expresión de genes, datos de información de empalme de ARN, datos de ARNip, datos de traducción de ARN y / o datos de actividad de proteínas. Del mismo modo, numerosos formatos de datos se consideran apropiados para su uso en este documento, sin embargo, los formatos de datos particularmente preferidos son los conjuntos de datos PARADIGM. El elemento del patrón determinante como se define en la reivindicación 1 puede variar considerablemente, sin embargo, los elementos del patrón determinante especialmente preferidos incluyen el estado de expresión de un gen, el nivel de proteína de una proteína y/o la actividad proteica de una proteína.

Según un segundo aspecto de la presente invención, se proporciona un sistema para el análisis *in sílico* de conjuntos de datos derivados de datos ómicos de células de acuerdo con la reivindicación 8. Normalmente, el sistema se programa además para generar datos de salida que comprenden una recomendación de un tratamiento para el paciente.

Tal y como se señaló anteriormente, también se contempla que al menos uno de los conjuntos de datos distintos se genera a partir de una muestra de un paciente que tiene una enfermedad neoplásica, y que se generan otros múltiples conjuntos de datos distintos a partir de cultivos celulares distintos que contienen células que no son del paciente. Preferiblemente, el paciente no ha sido tratado debido a la enfermedad neoplásica.

De acuerdo con un tercer aspecto de la presente invención, se proporciona un medio legible por ordenador [computadora] no transitorio que contiene instrucciones del programa de acuerdo con la reivindicación 10.

Más habitualmente, los datos ómnicos pueden incluir datos de número de copias de genes, datos de mutación genética, datos de metilación génica, datos de expresión génica, datos de información de empalme de ARN, datos de ARNip, datos de traducción de ARN y/o datos de actividad de proteínas, y se contempla especialmente que los conjuntos de datos distintos son conjuntos de datos PARADIGM.

Diversos objetos, características, aspectos y ventajas del ámbito de la invención se harán más evidentes a partir de la siguiente descripción detallada de las realizaciones preferidas, junto con las figuras de los dibujos que se acompañan en las que los números similares representan componentes similares.

Breve descripción de los dibujos

Las figuras 1A y 1B representan la sensibilidad de las líneas celulares de cáncer de mama frente a fármacos seleccionados (cisplatino 1A; geldanamicina 1B) en los paneles izquierdos, y representa esquemáticamente la actividad de los elementos del patrón en estas líneas celulares relacionadas con los fármacos seleccionados en los paneles derechos.

La figura 1C muestra la sensibilidad de una variedad de líneas celulares de cáncer de mama contra el cisplatino como se expresa en GI_{50} (panel superior) y el mapa de calor correspondiente para la expresión o regulación génica para las mismas células (panel inferior).

La figura 2A ilustra esquemáticamente un sistema de modelo de patrón en el que cada gen está representado a través de un modelo gráfico de factor estadístico.

La figura 2B representa esquemáticamente una modulación *in sílico* de un elemento del patrón de la figura 2A y los efectos descendentes asociados.

La figura 2C ilustra esquemáticamente una simulación de intervención farmacéutica en un ejemplo de sistema de modelado de patrones.

La figura 2D ilustra esquemáticamente el análisis de significación y la medición de desplazamiento de acuerdo con el ámbito de la invención.

La figura 3 ilustra esquemáticamente un experimento *in vivo* de validación para abatir en silicio un gen en una línea celular de cáncer de colon.

La figura 4 es una ilustración esquemática de un flujo de trabajo según el ámbito de la invención.

La figura 5A es un resultado ejemplar para los cambios pronosticados en la sensibilidad al cisplatino después de la manipulación *in sílico* de varias líneas celulares de cáncer en las que se eliminó IGFBP2.

La figura 5B es una salida ejemplar para los cambios pronosticados en la sensibilidad GSK923295 después de manipulación *in sílico* de varias líneas celulares de cáncer en las que se eliminó TP53INP1.

La figura 5C es un resultado ejemplar para los cambios pronosticados en la sensibilidad de fascaplisina después de manipulación *in sílico* de varias líneas celulares de cáncer en las que se eleminó ARHGEF25.

Descripción detallada

Basado en sistemas y métodos de análisis de patrones desarrollados recientemente como se describe con más detalle en los documentos WO 2011/139345, WO/2013/062505 y WO/2014/059036, los inventores ahora contemplan que el análisis de patrón y las modificaciones del modelo de patrón pueden usarse *in sílico* para identificar opciones de tratamiento farmacológico y/o simular elementos del patrón de tratamiento farmacológico que son determinantes o asociados con un parámetro relevante para el tratamiento (por ejemplo, resistencia a los medicamentos y/o sensibilidad a un tratamiento particular) de una afección, y especialmente una enfermedad neoplásica.

Más específicamente, los elementos del patrón identificados se modulan o modifican *in sílico* utilizando un sistema de análisis de patrones y un método para probar si se puede lograr el efecto deseado. Por ejemplo, donde un modelo de patrón para la resistencia a los medicamentos identifica la sobreexpresión de un determinado elemento como crítico para el desarrollo de una afección (por ejemplo, resistencia a los medicamentos contra un medicamento en particular), el nivel de expresión de ese elemento podría reducirse *in sílico* para probar de ese modo en el mismo sistema y método de análisis de patrones si la reducción de ese elemento *in sílico* potencialmente podría revertir la célula a la sensibilidad a los medicamentos. Tal enfoque es particularmente valioso cuando ya están disponibles múltiples líneas

celulares que representan múltiples variantes tumorales posibles. En tal caso, se puede realizar un análisis de patrón para cada una de las líneas celulares para obtener una colección de modelos de patrón específicos de línea celular. Tal recopilación es particularmente útil para la comparación con los datos obtenidos de una muestra del paciente, ya que los datos de la muestra del paciente pueden analizarse dentro del mismo espacio de datos que la recopilación, lo que finalmente permite la identificación de los objetivos de tratamiento para el paciente. Entre otras ventajas, los sistemas y métodos contemplados permiten, por lo tanto, el análisis de los datos del paciente de una muestra tumoral para identificar el tratamiento con múltiples medicamentos antes de que el paciente haya sido sometido al tratamiento farmacológico.

De esta forma, y visto desde una perspectiva diferente, los inventores han descubierto que se pueden usar varios datos ómicos de células y/o tejidos enfermos de un paciente en un enfoque computacional para determinar un perfil de sensibilidad para las células y/o tejidos, en donde el perfil está basado en identificación a priori de patrones y/o elementos de patrón en una variedad de células con enfermedades similares (por ejemplo, células de cáncer de mama). Más preferiblemente, el los patrones y/o los elementos del patrón identificados a priori están asociados con la resistencia y/o sensibilidad a una intervención farmacéutica particular y/o un régimen de tratamiento. Una vez que se establece el perfil de sensibilidad, el tratamiento puede predecirse directamente a partir de los patrones y/o los elementos de la patrón identificados a priori, o los patrones y/o los elementos del patrón identificados pueden ser modulados *in sílico* utilizando métodos y sistemas de modelado de patrones conocidos para ayudar a predecir resultados probables para la intervención farmacéutica y/o el régimen de tratamiento.

Cabe señalar que cualquier terminología que señale a un ordenador o computador debe entenderse que incluye cualquier combinación adecuada de dispositivos informáticos, incluidos servidores, interfaces, sistemas, bases de datos, agentes, pares, motores, controladores u otros tipos de dispositivos informáticos que funcionan de forma individual o colectiva. Se debería apreciar que los dispositivos informáticos comprenden un procesador configurado para ejecutar instrucciones de software almacenadas en un medio de almacenamiento tangible, no transitorio, legible por computador (por ejemplo, disco duro, unidad de estado sólido, RAM, flash, ROM, etc.). Las instrucciones de software configuran preferiblemente el dispositivo informático para proporcionar los roles, responsabilidades u otra funcionalidad como se discute a continuación con respecto al aparato divulgado. En realizaciones especialmente preferidas, los diversos servidores, sistemas, bases de datos o interfaces intercambian datos utilizando protocolos o algoritmos estandarizados, posiblemente basados en HTTP, HTTPS, AES, intercambios de claves público-privadas, API de servicios web, protocolos conocidos de transacciones financieras u otros métodos electrónicos para el intercambio de información. Los intercambios de datos se realizan preferiblemente a través de una red de paquetes conmutados, Internet, LAN, WAN, VPN u otro tipo de red de paquetes conmutados.

La mayoría de los pacientes con cáncer rara vez están sujetos a la monoterapia, sin embargo, la predicción precisa de una respuesta a combinaciones particulares de fármacos es uno de los desafíos más profundos en la terapia contra el cáncer. Dado que el número de posibles combinaciones de medicamentos es muy amplio, actualmente hay pocos datos estadísticamente significativos para respaldar cualquier combinación dada para un cáncer específico. En cambio, la mayoría de las terapias combinadas actuales se seleccionan a mano para apuntar a patrones independientes. Desafortunadamente, aunque los métodos actuales para diseñar terapias combinadas son algo pragmáticos, tienden a ser superficiales ya que no existe un enfoque estadístico preciso para identificar fármacos que sean candidatos para una terapia dual sinérgica. Además, la combinación numérica de predicciones de monoterapia no predecirá con precisión los resultados de las combinaciones, ya que los mecanismos de respuesta al fármaco no son necesariamente independientes.

Para abordar esta deficiencia, los inventores han desarrollado sistemas y métodos que incorporan el patrón de aprendizaje informado con predictores de monoterapia. Como se analiza con más detalle a continuación, generalmente se prefiere que se usen sistemas de modelado de patrón conocidos (preferiblemente PARADIGM) para inferir actividades de patrón a partir de datos de múltiples líneas celulares de células resistentes al tratamiento y sensibles al tratamiento (del mismo tipo de tumor). Así, los datos de actividad del patrón desarrollados se utilizan para construir modelos predictivos de respuesta a medicamentos en un enfoque como también se analiza con más detalle a continuación (topmodel [supermodelo]), y se inspecciona el modelo predictivo superior para cada medicamento para determinar qué genes a menudo tienen una alta ponderación de resistencia. Esos genes son entonces fijados *in sílico* en una posición de apagado en los sistemas de modelado de patrones conocidos (preferiblemente PARADIGM), y las actividades se vuelven a inferir, lo que en efecto simula *in sílico* el efecto anticipado de una intervención farmacológica *in vivo*. El supermodelo se utiliza para reevaluar los datos posteriores a la intervención recién inferidos. Como se puede apreciar fácilmente, donde la reevaluación indica un cambio de una predicción de resistencia a los medicamentos a una predicción de sensibilidad a los medicamentos, la intervención simulada *in sílico* puede traducirse en una recomendación de tratamiento para *in vivo* tratamiento.

A continuación, los inventores han demostrado la viabilidad de tales sistemas y métodos utilizando datos de líneas celulares de cáncer de mama conocidos y un gran panel de perfiles de respuesta a fármacos en monoterapia para estas células. Con el fin de simular el efecto de las terapias duales, los inventores

utilizaron los modelos de respuesta al fármaco altamente precisos capacitados sobre los datos del sistema de modelado del patrón como se describe más adelante, e inspeccionaron estos modelos basados en el sistema de modelado del patrón para candidatos genéticos que supuestamente se asociaron con la resistencia. Estas características asociadas a la resistencia fueron silenciadas *in sílico* en el sistema de modelado de patrones como un proxy para simular el efecto de una intervención farmacológica dirigida contra la acción de esos genes. Los modelos así obtenidos se utilizaron para reevaluar el conjunto de datos posterior a la intervención para un cambio hacia la sensibilidad. Si se observa un cambio, la inferencia es que la respuesta al medicamento que el modelo predijo *in sílico* probablemente será mejorado *in vivo* combinando un primer fármaco con un segundo tratamiento farmacológico objetivo basado en la lógica contra el gen candidato.

Debe apreciarse que la predicción del efecto de una combinación fármaco / característica-KO en este método requiere clasificadores lineales altamente precisos. Más preferiblemente, tales clasificadores usan datos del sistema de modelado de patrones (preferiblemente datos PARADIGM) como entrada para permitir su aplicación sin manipulación a los datos previos y posteriores a la intervención. Además, los modelos lineales también permitirán la inspección de los coeficientes de características para seleccionar características asociadas a la resistencia para simular la intervención que se trata.

Creación de modelos de predictores de respuesta a fármacos: los modelos predictivos promovidos para su uso en un entorno clínico deben tener un alto rendimiento. Para desarrollar un modelo predictivo de este tipo, generalmente se generan muchos modelos competitivos. Es necesario comparar el rendimiento de estos modelos competitivos múltiples para seleccionar a los mejores, pero los métodos para comparar estos resultados a menudo no son satisfactorios: por lo general, los parámetros entre las comparaciones varían tanto que no tienen sentido. Algunas herramientas de comparación de aprendizaje automático se han desarrollado para gestionar los parámetros de control. Por ejemplo, software como 'scikit-learn' y 'WEKA' están diseñados para recopilar con mucha precisión predicciones teóricas. Sin embargo, para disminuir el tiempo de ejecución, este software solo contiene temporalmente representaciones mínimas de datos en la memoria volátil. Por su diseño, se debe implementar un nuevo algoritmo predictivo dentro de su software para agregarlo a la comparación. Esto a menudo requiere traducir laboriosamente el código existente al lenguaje del código de canalización de aprendizaje automático (python para scikit-learn y Java para WEKA). Las comparaciones con algoritmos desarrollados fuera de estas herramientas de software siguen siendo extremadamente difíciles.

Para superar al menos algunas de estas dificultades, los inventores han desarrollado una herramienta ("topmodel" [o supermodelo]) que desacopla la gestión de datos de los algoritmos de aprendizaje automático aplicados a esos datos, lo que proporciona una línea flexible y de alto rendimiento. El topmodel [o supermodelo] lee datos, realiza particiones de capacitación y validación, realiza todas las transformaciones de datos y metadatos, y luego escribe esos datos en los diversos formatos requeridos por paquetes de software dispares. De esta manera, los mismos datos de capacitación y validación se exponen a diferentes algoritmos implementados en diferentes idiomas. El topmodel luego recopila resultados y los muestra en un formato unificado. En resumen, topmodel recopila datos accediendo a datos almacenados en cualquiera de los formatos de almacenamiento comunes (localmente o en servicios de almacenamiento en la nube), luego realiza una etapa de preprocesamiento en la que los datos y metadatos se someten a un preprocesamiento multiproceso, y en el que los datos se escriben en formatos de archivo requeridos por paquetes individuales de aprendizaje automático. Cabe señalar que este preprocesamiento es consistente entre formatos y se siembra (y, por lo tanto, es reproducible). En otra etapa más, se realiza la capacitación y la evaluación, y cada clasificador recibe capacitación sobre los datos de capacitación y se evalúa según los datos de validación. Lo anterior se realiza preferiblemente en un clúster, aumentando sustancialmente el rendimiento. Además de los modelos de evaluación, se construye un modelo totalmente capacitado sobre todo el conjunto de datos de entrada. En otra etapa de almacenamiento y visualización, se evalúa cada algoritmo y sus parámetros, y esas evaluaciones se recopilan en un formato de archivo unificado que puede almacenarse en una base de datos (consultable desde una interfaz de usuario). Por último, la interfaz define funciones para ejecutar modelos totalmente capacitados en datos novedosos, los usuarios pueden cargar sus datos a través de la interfaz y recibir predicciones.

Con respecto al etapa de recopilación de datos, se observa que para construir modelos predictivos, se deben recopilar conjuntos de datos de alta calidad con sus metadatos asociados. Hay muchas colecciones de datos de microarrays en dominio público. Sitios como el Gene Expression Omnibus (GEO) se han convertido en el depósito de intercambio de datos de facto para cientos de grandes grupos de población [cohortes] con los metadatos asociados necesarios. También hay consorcios generadores de datos a gran escala como SU2C y TCGA que brindan sus propios servicios de intercambio de datos. Sin embargo, debe reconocerse que la recopilación de estos conjuntos de datos requiere un esfuerzo significativo ya que cada sitio de almacenamiento tiene su propio sistema de consulta, formatos de archivo, políticas de uso, etc. Estos sistemas se actualizan constantemente. El acceso programático a estos conjuntos de datos directamente es extremadamente frágil. Por lo tanto, y en lugar de acceder directamente a estos repositorios para compartir datos, topmodel está configurado para leer datos y metadatos de cualquiera de los formatos de uso común. Esto incluye leer archivos delimitados por tabulaciones, archivos BED, acceder

a bases de datos mySQL y leer bases de datos SQLite. Además, la biblioteca topmodel C puede acceder tanto a bases de datos alojadas localmente como a bases de datos alojadas de forma remota.

Con respecto al preprocesamiento de datos, se observa que para que las comparaciones de rendimiento del modelo sean proporcionales, los datos expuestos a los paquetes de aprendizaje automático para capacitación deben ser consistentes. Para garantizar que los datos sean consistentes, topmodel ejecuta todos los preprocesamientos de datos antes de exponerlos a paquetes de aprendizaje automático. El preprocesamiento de datos incluye selección de características, transformaciones de datos y transformaciones de metadatos, y división en conjuntos de datos de capacitación y validación. Tal y como se debe apreciar, la selección de características es una estrategia común para aumentar la robustez. La reducción del espacio de características de entrada puede aliviar la 'maldición de la dimensionalidad' en la que se modela el ruido en lugar de la señal. La selección de características (en oposición a la reducción de características) es específicamente la eliminación de características menos informativas de los conjuntos de datos actuales. La implementación actual de topmodel admite el filtrado por variación mínima, rango de variación, relación mínima de ganancia de información y rango de ganancia de información. Además, los inventores reconocieron que transformar los datos en un espacio que aumenta la variación entre los subgrupos de interés puede aumentar el rendimiento de la predicción. Las transformaciones de datos que se convierten en un nuevo espacio de características se realizan preferiblemente antes de ingresar a topmodel para permitir el seguimiento de las características. Sin embargo, topmodel admite muchas transformaciones de datos que retienen el espacio de características de los conjuntos de datos originales: discretización por signo, rangos, umbrales de importancia y por expresiones booleanas.

Tal y como se reconocerá fácilmente, hay muchas formas de interpretar las variables de respuesta clínica. La interpretación de las variables de respuesta clínica es especialmente pertinente cuando se convierten variables continuas, como los datos IC50 en datos binarios (respondedores versus no respondedores) para su uso en algoritmos de clasificación binarios: múltiples umbrales diferentes para la división pueden ser opciones igualmente racionales. Por lo tanto, topmodel está configurado para admitir muchos esquemas de discretización de metadatos, incluso dividiéndolos alrededor de la mediana, por cuartiles superior e inferior, por signo, por rangos, por umbrales definidos por el usuario y por expresiones booleanas. Existen muchas técnicas para validar la solidez de la predicción. Además, las diferentes tareas de predicción deberían utilizar diferentes mediciones de robustez. Por ejemplo, LOOCV es más apropiado para cohortes muy pequeñas que RRS. Por lo tanto, topmodel también está configurado para admitir muchos métodos de validación diferentes. La técnica utilizada para medir la robustez se considera un parámetro en la línea de topmodel.

Cuando se toman en combinación, las opciones en origen de datos, selección de características de datos, transformación de datos y transformación de metadatos, y método de validación, describen un gran espacio potencial de entradas. El tiempo de procesamiento y las necesidades de almacenamiento para estas etapas de preprocesamiento son importantes y, por lo tanto, topmodel requiere un gran sistema de almacenamiento accesible para un clúster informático. Topmodel envía archivos de capacitación y validación a un sistema de almacenamiento de enjambre, que es de gran capacidad y redundante. El enjambre también está ensamblado para ser accesible a los clústeres informáticos, lo que hace que estos archivos estén directamente disponibles para capacitación. El topmodel utiliza varias técnicas para reducir el tiempo de preprocesamiento. En lugar de descargar el conjunto de datos cada vez para cada modelo, topmodel descarga datos una vez y los guarda en la memoria. Se utilizan copias internas de los datos para realizar la selección y transformación de características. Estas etapas de manipulación de datos están encadenados para que no se repita ningún trabajo. Además, los módulos de preprocesamiento de topmodel son multiproceso. El enhebrado permite que las etapas de preprocesamiento se ejecuten simultáneamente, ahorrando tiempo y al mismo tiempo compartiendo memoria, lo que puede ayudar a evitar la repetición del trabajo.

El preprocesamiento aumenta exponencialmente con el número de parámetros que se exploran. Al explorar múltiples conjuntos de datos con múltiples métodos de selección de características y múltiples transformaciones de datos, el preprocesamiento puede convertirse en el cuello de botella en la línea del modelo superior. El enfoque actual de subprocesos múltiples puede generar miles de manipulaciones únicas de conjuntos de datos en unas pocas horas.

Con respecto a la capacitación y la evaluación, debe apreciarse que topmodel utiliza comandos muy simples de 'capacitación' y 'clasificación' para construir y probar modelos, y que todos los paquetes de aprendizaje automático en topmodel se ejecutan desde un comando similar a UNIX. Los paquetes compatibles deben tener dos ejecutables: un comando de capacitación y un comando de clasificación. El comando de capacitación debe recibir como entrada al menos un archivo de datos y generar al menos un archivo de modelo. El comando de clasificación debe recibir como entrada al menos un archivo de datos y un archivo de modelo y generar al menos un archivo de resultados. Este es un esquema muy común para algoritmos de aprendizaje automático que se admite fácilmente. Por ejemplo, los ejecutables 'capacitar y 'clasificar' salen de la caja para svm-light. Para otros algoritmos que no se ejecutan desde la línea de comandos de esta manera, los inventores desarrollaron pequeños contenedores. Por ejemplo, los modelos

glmnet (es decir, regression contraída *[ridge regression]*, lasso y las redes elásticas) se ejecutan típicamente desde dentro de R, por lo que no tienen una interfaz de línea de comandos. Los inventores desarrollaron dos pequeños módulos R, uno para capacitación y otro para clasificación, que pueden ejecutarse desde la línea de comando usando R en modo por lotes.

Modelos de capacitación: Los modelos de capacitación son la etapa más costosa desde el punto de vista computacional en la línea de topmodel. La capacitación de modelos complejos (por ejemplo, máquinas de vectores de soporte de kernel polinomiales) sobre un conjunto de datos con miles de características puede llevar horas en completarse en nuestros nodos de clúster enjambre (procesadores Intel Xeon de cuatro núcleos). Hay al menos dos trabajos de capacitación por modelo en topmodel: Un conjunto de trabajos de capacitación para evaluar el desempeño (por ejemplo, modelos de validación cruzada) y un modelo totalmente capacitado que utiliza todo el conjunto de datos como entrada. Debido a la etapa de preprocesamiento, los modelos de capacitación pueden ser completamente paralelos. Todos los modelos están capacitados en nodos independientes en nuestro sistema de clúster. Al dividir estos trabajos de capacitación, el tiempo necesario para generar miles de modelos está limitado principalmente por el tamaño del clúster.

Clasificación: hay al menos tres trabajos de clasificación por modelo en topmodel: un conjunto de trabajos de clasificación para evaluación en el conjunto de datos de validación, un conjunto de trabajos de clasificación para volver a inspeccionar el conjunto de datos de capacitación y un trabajo de clasificación para inspeccionar el modelo completamente capacitado. De manera similar a la capacitación, todas las etapas de clasificación se pueden ejecutar en paralelo en el clúster (después de que la capacitación haya finalizado). La clasificación utiliza relativamente pocos recursos informáticos en comparación con la capacitación.

Modelos de evaluación: después de completar toda la clasificación, un módulo en topmodel lee los archivos de resultados generados por paquetes dispares de aprendizaje automático y convierte esa información en un formato de informe unificado. Se genera un archivo de informe por modelo y se almacena en el enjambre. Como se trata de una etapa por modelo, también se puede ejecutar en el clúster. Este formato de informe describe qué muestras se usaron en la capacitación, que fueron los puntajes de predicción sin procesar del algoritmo de clasificación y cuál fue la precisión de las predicciones tanto en la capacitación como en los cohortes de prueba. Para los modelos lineales, este formato también incluye hasta 200 nombres de genes y sus coeficientes en el modelo predictivo.

Almacenamiento de resultados: después de completar todas las evaluaciones, un módulo en topmodel reúne todos los resultados en un solo archivo de informe unificado. Este archivo describe todas las tareas de predicción, métodos de selección de características, transformaciones de datos, subgrupos de metadatos y estadísticas del modelo. El módulo topmodel que reúne estos resultados verifica que cada entrada sea única, asegurando que no haya duplicación en los resultados. Este archivo de informe actúa como una base de datos basada en archivos de resultados de topmodel. En un aspecto preferido, otro módulo en topmodel refleja estos resultados de topmodel en una base de datos que se puede consultar desde la web. Luego se proporciona una interfaz de usuario que permite visualizar los resultados consultados desde la base de datos.

Predicción con topmodel: se pueden usar modelos totalmente capacitados para predecir datos nuevos enviados por el usuario. Usando la interfaz de usuario topmodel, los usuarios pueden cargar datos delimitados por tabulaciones para sus muestras. El CGI de topmodel guarda sus datos en el espacio temporal local. Luego hace coincidir las características de los datos del usuario con el modelo que se solicita. Cuando faltan valores en los datos del usuario, se insertan valores nulos. El modelo solicitado se utiliza para calificar los datos del usuario utilizando un módulo en la biblioteca de topmodel C. Los puntajes se informan a la interfaz de usuario de topmodel en formato JSON, y los datos del usuario se borran del disco. La interfaz de usuario de topmodel recibe los puntajes de predicción en formato JSON y los representa en un gráfico. En este gráfico se incluye un gráfico circular que muestra la superposición de características entre los datos enviados por el usuario y el modelo que se aplica. Además, los puntajes de predicción del conjunto de datos de capacitación también se trazan para dar contexto a partir de ejemplos verdaderamente positivos y verdaderos negativos.

Debe apreciarse que los sistemas y métodos también serán adecuados para la identificación del mecanismo de acción y/u objetivo de un nuevo compuesto terapéutico. Por ejemplo, múltiples y distintas células y/o tejidos (normalmente células o tejidos enfermos) se exponen a uno o más compuestos candidatos para evaluar un posible efecto terapéutico. Por lo general, dicho efecto se medirá como un IG50, IC50, inducción de apoptosis, cambio fenotípico, etc., para cada una de las células y/o tejidos múltiples y distintos, y el aprendizaje automático tal y como se describe en el presente documento, se emplea para identificar uno o más elementos determinantes del patrón en los conjuntos de datos de las células y/o tejidos. Tal identificación conducirá fácilmente a un objetivo potencial y/o mecanismo de acción para el nuevo compuesto terapéutico. Además, los sistemas y métodos contemplados también serán adecuados para identificar fármacos secundarios (por ejemplo, fármacos quimioterapéuticos conocidos) que pueden

aumentar la eficacia del nuevo compuesto terapéutico. En consecuencia, utilizando los sistemas y métodos descritos en este documento, debe reconocerse que el modo de acción y los objetivos moleculares pueden identificarse para un nuevo fármaco, así como también pueden identificarse combinaciones sinérgicas de nuevos fármacos o fármacos conocidos.

De la misma manera, también debe reconocerse que pueden identificarse nuevos objetivos para un medicamento existente para el que no existe ningún compuesto farmacéutico. Por ejemplo, cuando los sistemas y métodos presentados en este documento indican un elemento de patrón particular como un elemento de patrón determinante para un tratamiento exitoso para el cual no existe un medicamento actual, se puede emplear un diseño de medicamento racional para desarrollar conductores e incluso compuestos farmacéuticos activos (por ejemplo, anticuerpos, inhibidores enzimáticos, etc.) que se dirigen específicamente a estos elementos determinantes de la patrón tan identificados.

Por lo tanto, los inventores también contemplan el método de análisis *in sílico* de conjuntos de datos derivados de datos ómicos de células de acuerdo con la reivindicación 1 para la identificación de un objetivo farmacológico y/o mecanismo de acción. Tales métodos incluirán habitualmente una etapa de acoplamiento informativo de una base de datos del modelo de patrón a un sistema de aprendizaje automático y un motor de análisis de patrón, en el que la base de datos del modelo de patrón almacena conjuntos de datos múltiples y distintos derivados de datos ómicos de células múltiples y distintas tratadas con un compuesto candidato (por ejemplo, fármaco quimioterapéutico, anticuerpo, inhibidor de quinasa, etc.), respectivamente, y en el que cada conjunto de datos comprende una pluralidad de datos de elementos del patrón. Posteriormente, un sistema de aprendizaje automático recibirá los distintos conjuntos de datos, y el sistema de aprendizaje automático identificará un elemento de patrón determinante en los distintos conjuntos de datos que está asociado con la administración del compuesto candidato a las células sustancialmente como se describe en este documento. En otra etapa, el motor de análisis de patrón recibirá al menos uno de los conjuntos de datos distintos de las células y asociará el elemento de patrón determinante en el conjunto de datos distinto con un patrón específico o un objetivo farmacológico. El patrón específico identificado o el objetivo farmacológico se usa a continuación en una salida (por ejemplo, archivo de informe opcionalmente con representación gráfica) que correlaciona el compuesto candidato con el patrón específico o el objetivo farmacológico. El motor de análisis de patrón se usa para modular el elemento de patrón determinante recientemente identificado en el conjunto de datos para producir un conjunto de datos modificado desde la célula, y el sistema de aprendizaje automático puede identificar (en base al conjunto de datos modificado) un cambio en el estado de un parámetro de tratamiento para la célula.

Ejemplos

Tal y como es bien conocido, diferentes líneas celulares de un tejido enfermo (por ejemplo, de cáncer de mama) tienen una expresión y un entorno regulador muy diferentes en respuesta al tratamiento con un medicamento en particular. Por ejemplo, mientras que algunos tipos de cáncer de mama (por ejemplo, basal, no basal) tendrán una sensibilidad distinta hacia el cisplatino, como se muestra en el gráfico de la figura 1A, otros tipos de cáncer de mama (ERBB2AMP, no ERBB2AMP) tendrán una sensibilidad distinta hacia la geldanamicina tal y como se muestra en la gráfica de figura 1B. Las ilustraciones esquemáticas correspondientes para las figuras 1A y B ubicados a la derecha de los gráficos ilustran la información del patrón ejemplar correspondiente para las respectivas células / tratamientos farmacológicos donde las líneas continuas indican la activación de la transcripción, las líneas discontinuas representan la activación de la quinasa y una barra al final de una línea representa el efecto inhibitorio.

El panel superior de la figura 1C representa una vista más detallada de la sensibilidad a los medicamentos de varias líneas celulares de cáncer de mama contra cisplatino, mientras que el panel inferior muestra un mapa de calor de expresión / regulación en las mismas líneas celulares (indicadas en el eje x) con respecto a varios elementos objetivo (indicados en el eje y, vea también la ilustración esquemática de la figura 1A) dentro de un patrón de la célula cancerosa. Tal y como se puede reconocer fácilmente, la expresión y la regulación génica son sustancialmente diferentes de una línea celular a otra, sin un patrón aparente asociado con la sensibilidad o resistencia al cisplatino. Por lo tanto, si bien existe una gran cantidad de información genómica, un experto en la materia carece de orientación efectiva o incluso informativa de estos datos para identificar una estrategia o recomendación de tratamiento adecuada.

Para el presente ejemplo, se usó un panel de 50 líneas celulares de cáncer de mama para proporcionar un conjunto de datos adecuado para demostrar la efectividad de los sistemas y métodos (topmodel) contemplados aquí. Además de tener datos de varios ensayos de todo el genoma, se ha analizado la respuesta a 138 fármacos en estas líneas celulares. Como resultado, se pueden analizar muchos desafíos de predicción en este conjunto de datos mientras se mantiene el constante efecto de cohorte. Más específicamente, se obtuvieron datos de expresión de microarrays Exon de Affymetrix y el número de copia de microarrays SNP 6.0 Genome Wide de Affymetrix para 50 líneas celulares de cáncer de mama y estos datos se usaron para inferir actividades del patrón usando sistemas de modelado de patrón conocidos

(como se describe en WO 2011/139345 y WO 2013/062505). Los datos que resultan de dicha transformación de los datos de expresión y número de copias son una matriz de características de patrón por muestras apropiadas para su uso en sistemas y métodos (topmodel) contemplados en este documento. Además de los datos genómicos, se obtuvieron datos de respuesta a fármacos IC50 (GI_{50}, Amax, ACarea, ACarea filtrada y dosis máxima) para 138 fármacos.

Estos datos se usaron para construir clasificadores de respuesta a medicamentos (sensibles frente a resistentes) en la línea de topmodel como se describe en la siguiente tabla a continuación. En combinación, estos parámetros describen uno posibles 129,168 modelos totalmente capacitados. Como cada modelo se valida con una validación cruzada de 5x3 veces, esto requiere la capacitación de otros 15 modelos por modelo totalmente capacitado, o 1,937,520 modelos de evaluación adicionales. El número total de modelos a capacitar supera los 2 millones.

Conjuntos de datos	Exón expresión, número de copia SNP6, PARADIGM
Meta conjuntos de datos	138 respuesta a fármacos IC50
Subgrupos	mediana IC50, mediana GI_{50}, mediana Amax, mediana ACarea, mediana ACarea filtrada, mediana dosis máxima
Clasificadores	NMFpredictor, SVMlight (kernel lineal), SVMlight (kernel polinomial de primer orden), SVMlight (kernel polinomial de segundo orden), WEKA SMO, WEKA j48 árboles, hiperpipes WEKA, bosques aleatorios WEKA, Bayes ingenuos WEKA, reglas WEKA JRip, glmnet lasso, glmnet regresión contraída, redes elásticas glmnet
Métodos de selección de funciones	Ninguna, clasificación de varianza (20 características), clasificación de varianza (200 características), clasificación de varianza (2000 características)
Método de validación	5x3 veces validación cruzada

Para los datos de la línea celular de cáncer de mama mencionados anteriormente, se seleccionó el modelo lineal más preciso para cada fármaco (de los 138 fármacos disponibles) para su posterior análisis, y para cada modelo se extrajeron hasta 200 características asociadas a la resistencia mediante la inspección de los coeficientes en estos modelos lineales y reportando las características de mayor clasificación. De las 17.325 características en los patrones, 5.065 fueron seleccionadas por al menos uno de los 138 modelos de respuesta a fármacos como asociadas con resistencia. De estas 5.065 características, las 200 que estaban asociadas con la resistencia con mayor frecuencia fueron seleccionadas para eliminar *in sílico*.

Modulación de patrón *in sílico*: sistemas de modelado de patrón preferidos tal y como se describe en WO 2011/139345, WO 2013/062505 y WO 2014/059036 enseñan sobre las actividades del patrón inferida ajustando los datos biológicos observados (datos ómnicos) a un módulo de dogma central (generalmente basado en datos curados a priori de información de patrón conocida), permitiendo que muchos módulos se propaguen entre sí hasta que converjan en un estado estable. La figura 2A proporciona una ilustración esquemática de un modelo de patrón (PARADIGM) en el que un gen se representa mediante un modelo gráfico de factor estadístico.

Como debería apreciarse fácilmente, tales sistemas de modelado de patrones también pueden usarse para simular el efecto de una intervención dirigida. Por ejemplo, como se ilustra esquemáticamente en figura 2B para el silenciamiento genético de un gen, el nodo de ARNm objetivo en el módulo dogma central puede ser forzado a un estado suprimido, y las actividades del patrón se pueden volver a inferir. Además, el nodo de ARNm abatido se puede desconectar de sus nodos principales, lo que inhibirá el estado bajo de ARNm que se propaga espuriosamente su estado suprimido a los reguladores transcripcionales del gen objetivo. Un ejemplo esquemático adicional se proporciona en la figura 2C en donde, en el panel (a), una patrón ejemplar se expresa como un gráfico de factores que permite ventajosamente modelar e inferir actividades del patrón. Los nodos de evidencia se pueblan utilizando datos que se derivan de ensayos de todo el genoma (normalmente datos ómicos) como datos de expresión y datos de números de copias. Por lo tanto, las señales de estos nodos se propagan a través del gráfico de factores. El panel (b) muestra esquemáticamente una simulación de intervención. En la característica objetivo (desactivación de la expresión génica), los nodos de evidencia se desconectan y el nodo de ARNm se sujeta a un estado regulado negativamente.

Usando el sistema anterior, se realizaron simulaciones de intervención para las 200 características

asociadas a la resistencia en las líneas celulares de cáncer de mama, lo que genera 200 nuevos conjuntos de datos 'posteriores a la intervención', cada uno de los cuales representa el efecto de un silenciamiento genético dirigido. Para cuantificar el efecto de las intervenciones duales, se aplica un modelo de respuesta al fármaco a los conjuntos de datos previos y posteriores a la intervención y se observa el cambio en la resistencia prevista. La magnitud de este cambio indica cuánto sinergiza la intervención característica con la respuesta de monoterapia que predice el modelo.

Análisis de significación y medición de desplazamiento: El siguiente análisis de significación se realizó para ajustar aún más los resultados. En el ejemplo anterior de cáncer de mama, cada modelo lineal seleccionado para el análisis podría designar 200 características como asociadas a la resistencia. Como solo los 200 principales fueron seleccionados de la lista completa de más de 5,000 nominados, cada modelo lineal contenía ciertas características que fueron seleccionadas y otras características que no fueron seleccionadas. En promedio, un modelo lineal dado tiene 3 características en el conjunto asociado a la resistencia 200. Por lo tanto, para cualquier modelo de respuesta dado hay un conjunto de alrededor de 197 conjuntos abatidos de datos simulados que no están relacionados con el modelo, que se utilizan para crear una distribución empírica nula. Los mejores modelos para cada medicamento se aplican a todos los conjuntos de datos de abatimiento de características, y aquellos que no están relacionados con el medicamento que se analiza crean un modelo de fondo con el que medir la importancia de cada gen que se seleccionó, tal y como se ilustra esquemáticamente en la figura 2D. En este punto, el panel (a) ilustra esquemáticamente los modelos de respuesta al fármaco A, B y C, cada uno de los cuales contiene hasta 200 genes previamente identificados como relacionados con la resistencia, y algunos de los genes entre los modelos A, B y C pueden superponerse. Al analizar las combinaciones de fármaco / característica-KO del modelo C, todos los genes, x, se usaron del conjunto $x \in \{A \cup B - C\}$, en un modelo nulo. En el panel (b), el modelo C se aplica a todos los genes $X \in \{A \cup B - C\}$ y todas las muestras $i \in N$. La cantidad de desplazamiento para cada combinación de KO / fármaco / muestra, $\Delta_{x,c,i}$ se graba en un modelo de fondo. El modelo C también se aplica a cada gen $y \in \{C\}$, y la cantidad de desplazamiento, $\Delta_{y,c,j}$ grabado. Como se muestra en el panel (c), la cantidad de desplazamiento en una combinación seleccionada de fármaco / gen / muestra se mide luego por su importancia frente a la distribución de fondo de genes no relacionados.

Para validar dicho enfoque conceptual, los inventores usaron la línea celular HT29 de cáncer de colon en un conjunto de experimentos como se muestra esquemáticamente en la figura 3. En un primer experimento *in vitro*, un ARNip contra GFP (proteína verde fluorescente) se expresó en la célula como control negativo (ya que las células HT29 no expresan GFP), mientras que en un segundo experimento *in vitro*, se expresó un ARNip contra GNAI3 para eliminar la expresión de GNAI3 nativa en la célula. Se obtuvieron datos de ómicos (número de copia del gen, nivel de expresión, datos de proteómica) para ambos experimentos *in vitro*, y el análisis del patrón se realizó con PARADIGM. En un experimento *in sílico* independiente, GNAI3 se ajustó artificialmente a 'no expresión', y se realizaron pruebas T emparejadas como se indica en la figura 3 para ver si las condiciones experimentales observadas en las células GNAI3-abatidas in vitro se correlacionarían más estrechamente con las ´células GNAI3-abatidas *in sílico* que con las células GFP-abatidas *in vitro*. Sorprendentemente, los resultados *in sílico* son paralelos a los resultados in vitro con un grado relativamente alto de significación estadística. Por lo tanto, la utilidad potencial del enfoque anterior se indicó claramente.

En vista de lo anterior, la figura 4 ilustra esquemáticamente una realización típica del ámbito de la invención tal como se presenta en el presente documento. Aquí, los datos ómicos (preferiblemente como conjuntos de datos PARADIGM) del mismo tipo celular pero diferente sensibilidad a los medicamentos (por ejemplo, sensible frente a resistente, como se expresa a través y sobre la base de los valores GI_{50}) se someten a análisis de aprendizaje automático en un centro de aprendizaje automático utilizando topmodel para identificar elementos del patrón putativos que confieren resistencia y/o sensibilidad hacia el fármaco tal y como se describió anteriormente. Una vez identificados, uno o más elementos putativos del patrón se modulan artificialmente *in sílico* (en este punto: como una abatida simulada), y los conjuntos de datos así obtenidos se someten a un análisis adicional para predecir si (y en qué grado) la modificación resultó en un cambio en la sensibilidad al fármaco. Los resultados del análisis se proporcionan luego en un formato de salida que permite la identificación de elementos del patrón que proporcionarán o contribuirán a un cambio deseado en la resistencia a los medicamentos. En el ejemplo de la figura 4, el cambio calculado / simulado en la sensibilidad contra cisplatino tras abatir el IGFBP2 en células de cáncer de mama se indica para cada línea celular usando flechas. Las figuras 5A-5C muestran los resultados pronosticados para los cambios en la sensibilidad a los medicamentos en función de un cambio calculado / simulado en la expresión de un elemento del patrón previamente identificado de las células de cáncer de mama. Más específicamente, la figura 5A representa la sensibilidad al cisplatino y el elemento del patrón es IGFB2, la figura 5B representa la sensibilidad GSK923295 y el elemento de patrón es TP53INP1, mientras que la figura 5C representa la sensibilidad a la fascaplisina y el elemento del patrón es ARHGEF25.

Por supuesto, se debe apreciar que los ejemplos anteriores solo proporcionan una ilustración del ámbito inventivo y no deben considerarse limitantes. De hecho, si bien los ejemplos solo proporcionan un análisis de la modulación de elementos de patrón individual, se debe apreciar que los elementos de patrón múltiple se pueden modificar, concurrentemente o secuencialmente. Más aún, debe reconocerse que si bien se

discuten los cambios de abatida, todas las modificaciones (por ejemplo arriba, abajo, [heterólogo o de otra manera recombinante] expresión de gen) se consideran adecuados para su uso en el presente documento. Tales modificaciones pueden ser modificaciones directas en el nivel de ácido nucleico (por ejemplo, abatir, eliminar, borrar, expresión mejorada, estabilidad mejorada, etc.) y/o en el nivel de proteína (por ejemplo, mediante anticuerpos, expresión recombinante, inyección, etc.) o modificaciones indirectas mediante componentes reguladores (por ejemplo, proporcionando estimuladores de expresión, represores de transcripción, etc.).

Todavía más, se debe tener en cuenta que, si bien los ejemplos anteriores se utilizan para interferir con un única patrón o red de patrones, *in sílico* e *in vivo*, también se contemplan manipulaciones que afectan a múltiples patrones, estén o no asociadas funcionalmente entre sí. Del mismo modo, debe reconocerse que la manipulación del patrón también se puede realizar de manera que se establezca artificialmente un resultado deseado, y que luego se realice un análisis posterior para identificar parámetros que puedan modificarse para conducir al resultado deseado. Además, aunque PARADIGM es un sistema de modelo de patrón particularmente preferido, debe apreciarse que todos los sistemas de modelado de patrón se consideran adecuados para su uso en el presente documento. Más habitualmente, estos sistemas de modelado tendrán al menos un componente conocido a priori.

De esta forma, se han descrito realizaciones y aplicaciones específicas de métodos de redes de respuesta a fármacos. Debería ser evidente para los expertos en la materia que hay muchas más posibles modificaciones además de las ya descritas sin apartarse de los conceptos inventivos del presente documento. El ámbito de la invención, por lo tanto, no debe restringirse excepto por las reivindicaciones adjuntas. Además, al interpretar tanto la especificación como las reivindicaciones, todos los términos deben interpretarse de la manera más amplia posible de acuerdo con el contexto. En particular, los términos "incluye" y "comprende" deben interpretarse como que se refieren a elementos, componentes o etapas de una manera no exclusiva, lo que indica que los elementos, componentes o etapas a los que se hace referencia pueden estar presentes, o utilizados, o combinados con otros elementos, componentes o etapas que no se mencionan expresamente. Cuando las reivindicaciones de especificación se refieren al menos a uno de algo seleccionado del grupo que consiste en A, B, C ... y N, el texto debe interpretarse como que requiere solo un elemento del grupo, no A más N o B más N, etc.

REIVINDICACIONES

1. Método de análisis *in sílico* de conjuntos de datos derivados de datos ómicos de células, que comprende:

acoplar informativamente una base de datos de modelo de patrón a un sistema de aprendizaje automático y a un motor de análisis de patrón;

en donde la base de datos del modelo de patrón almacena una pluralidad de conjuntos de datos ómicos que comprenden datos ómicos de una pluralidad de células enfermas distintas, respectivamente, y en donde cada conjunto de datos comprende una pluralidad de datos de elementos de patrón;

recibir, mediante el sistema de aprendizaje automático, la pluralidad de conjuntos de datos;

identificar, mediante el sistema de aprendizaje automático, un elemento determinante del patrón en la pluralidad de conjuntos de datos que está asociado con un estado de un parámetro de tratamiento de las células enfermas; el elemento determinante del patrón es una resistencia al tratamiento asociada o una sensibilidad al tratamiento asociada a los datos del patrón;

recibir, mediante el motor de análisis de patrón, al menos uno de los conjuntos de datos de las células enfermas;

modular *in sílico*, mediante el motor de análisis de patrón, el elemento de patrón determinante en al menos uno de los conjuntos de datos para producir un conjunto de datos modificado de la célula enferma, en el que el conjunto de datos modificado incluye al menos un elemento de patrón modificado y el al menos un elemento de patrón modificado se modifica directamente en un nivel de ácido nucleico o un nivel de proteína, o indirectamente a través de un componente regulador; y además en donde la modulación *in sílico* comprende:

- representar *in sílico* el modelo de patrón a través de un modelo gráfico de factores que comprende nodos de factores y nodos de evidencia variable; los nodos de evidencia variable se llenan utilizando los datos ómnicos derivados;
- forzar *in sílico* el nodo de evidencia variable que representa el elemento determinante del patrón del modelo de patrón en un estado suprimido; y
- volver a inferir *in sílico* las actividades del patrón para obtener el conjunto de datos modificado; e

identificar, mediante el sistema de aprendizaje automático y utilizar el conjunto de datos modificado, un cambio en el estado del parámetro de tratamiento para la célula enferma.

2. El método de la reivindicación 1, en el que al menos uno de los conjuntos de datos se genera a partir de la muestra de un paciente que tiene una enfermedad neoplásica, y en el que se generan otros muchos conjuntos de datos a partir de cultivos celulares distintos que contienen células que no son del paciente; preferiblemente,

en donde el paciente no ha sido tratado debido a la enfermedad neoplásica; o

que comprende además una etapa de generar datos de salida que comprenden una recomendación de tratamiento para el paciente.

3. El método de la reivindicación 1, en donde la pluralidad de células enfermas distintas difieren entre sí con respecto a la sensibilidad de las células a un fármaco; o

en donde un primer conjunto de la pluralidad de células enfermas distintas es sensible al tratamiento con un fármaco, y en donde un segundo conjunto de la pluralidad de células enfermas distintas es resistente al tratamiento con el fármaco.

4. El método de la reivindicación 1, que comprende además una etapa de identificación de un fármaco que tiene como objetivo el elemento determinante del patrón cuando el cambio en el estado del parámetro de tratamiento excede un umbral predeterminado.

5. El método de la reivindicación 1, en el que los datos ómicos se seleccionan del grupo que consiste en datos de número de copia génica, datos de mutación génica, datos de metilación génica, datos de expresión génica, datos de información de empalme de ARN, datos de ARNip, datos de mobilidad de ARN y

datos de actividad de proteínas.

6. El método de la reivindicación 1, en el que el cambio de estado es un cambio de resistencia al fármaco a sensibilidad al fármaco.

7. El método de la reivindicación 1, que comprende además una etapa de preprocesamiento de los conjuntos de datos que incluye selección de características, transformación de datos, transformación de metadatos y/o división en conjuntos de datos de capacitación y validación.

8. Sistema para el análisis *in sílico* de conjuntos de datos derivados de datos ómicos de células, que comprende:

una base de datos del modelo de patrón acoplada informativamente a un sistema de aprendizaje automático y un motor de análisis de patrón;

en donde la base de datos del modelo de patrón está programada para almacenar una pluralidad de conjuntos de datos ómicos que comprenden datos ómicos de una pluralidad de células enfermas distintas, respectivamente, y en donde cada conjunto de datos comprende una pluralidad de datos de elementos de patrón;

en el que el sistema de aprendizaje automático está programado para recibir de la base de datos del modelo de patrón la pluralidad de conjuntos de datos, y en el que el sistema de aprendizaje automático está además programado para identificar un elemento de patrón determinante en la pluralidad de conjuntos de datos que está asociado con el estado de un parámetro de tratamiento de las células enfermas; el elemento determinante del patrón es una resistencia al tratamiento asociada o un dato asociado a la sensibilidad del tratamiento;

en el que el motor de análisis de patrón está programado para recibir al menos uno de los conjuntos de datos de las células enfermas y además está programado para modular *in sílico* el elemento de patrón determinante en al menos uno de los conjuntos de datos para producir un conjunto de datos modificado de la célula enferma;

en el que el conjunto de datos modificado incluye al menos un elemento de patrón modificado y el al menos un elemento de patrón modificado se modifica directamente en un nivel de ácido nucleico o un nivel de proteína, o indirectamente a través de un componente regulador; y además en donde la modulación *in sílico* comprende:

- representar *in sílico* el modelo de patrón a través de un modelo gráfico de factores que comprende nodos de factores y nodos de evidencia variable; los nodos de evidencia variable se llenan utilizando los datos ómnicos derivados;
- forzar *in sílico* el nodo de evidencia variable que representa el elemento determinante del patrón del modelo de patrón en un estado suprimido; y
- volver a inferir *in sílico* las actividades del patrón para obtener el conjunto de datos modificado; y

en donde el sistema de aprendizaje automático está programado para identificar un cambio en el estado del parámetro de tratamiento para la célula enferma usando el conjunto de datos modificado.

9. El sistema de la reivindicación 8, en el que al menos uno de los conjuntos de datos se genera a partir de una muestra de un paciente que tiene una enfermedad neoplásica, y en el que se generan muchos otros conjuntos de datos a partir de cultivos celulares distintos que contienen células que no son del paciente; preferiblemente,

en donde el paciente no ha sido tratado debido a la enfermedad neoplásica; o

en donde el sistema de aprendizaje automático está programado para generar datos de salida que comprenden una recomendación de tratamiento para el paciente.

10. Medio no transitorio legible por ordenador [computador] que contiene instrucciones de programa para hacer que el sistema de la reivindicación 8 realice un método que comprende los etapas de:

transferir desde la base de datos del modelo de patrón al sistema de aprendizaje automático una pluralidad de conjuntos de datos ómicos que comprenden datos ómicos de una pluralidad de células enfermas distintas, respectivamente, y en donde cada conjunto de datos comprende una

pluralidad de datos de elementos del patrón;

identificar, mediante el sistema de aprendizaje automático, un elemento determinante del patrón en la pluralidad de conjuntos de datos que está asociado con un estado de un parámetro de tratamiento de las células enfermas; el elemento determinante del patrón es una resistencia asociada al tratamiento o un dato asociado a la sensibilidad del tratamiento;

recibir, mediante el motor de análisis de patrón, al menos uno de los conjuntos de datos de las células enfermas;

modular *in sílico*, mediante el motor de análisis de patrón, el elemento de patrón determinante en al menos uno de los conjuntos de datos para producir un conjunto de datos modificado de la célula enferma; en el que el conjunto de datos modificado incluye al menos un elemento de patrón modificado y al menos uno de los elementos de patrón modificado se modifica directamente en un nivel de ácido nucleico o un nivel de proteína, o indirectamente a través de un componente regulador; y además en donde la modulación *in sílico* comprende:

- representar *in sílico* el modelo de patrón a través de un modelo gráfico de factores que comprende nodos de factores y nodos de evidencia variable; los nodos de evidencia variable se llenan utilizando los datos ómnicos derivados;
- forzar *in sílico* el nodo de evidencia variable que representa el elemento determinante del patrón del modelo de patrón en un estado suprimido; y
- volver a inferir *in sílico* las actividades del patrón para obtener el conjunto de datos modificado; e

identificar, mediante el sistema de aprendizaje automático y utilizar el conjunto de datos modificado, un cambio en el estado del parámetro de tratamiento para la célula enferma.

11. El medio no transitorio legible por ordenador [computador] de la reivindicación 10, en el que los datos ómicos se seleccionan del grupo que consiste en datos de número de copia génica, datos de mutación génica, datos de metilación génica, datos de expresión génica, datos de información de empalme de ARN, datos de ARNip, datos de traducción de ARN, y datos de actividad proteica.

12. El método de la reivindicación 1, que comprende además

asociar, mediante el motor de análisis del patrón, el elemento determinante del patrón en al menos uno de los conjuntos de datos distintos con un patrón específico o un objetivo farmacológico, y producir una salida que correlaciona el compuesto candidato con el patrón específico o el objetivo farmacológico; preferiblemente,

en donde el compuesto candidato es un fármaco quimioterapéutico.

FIG. 1A

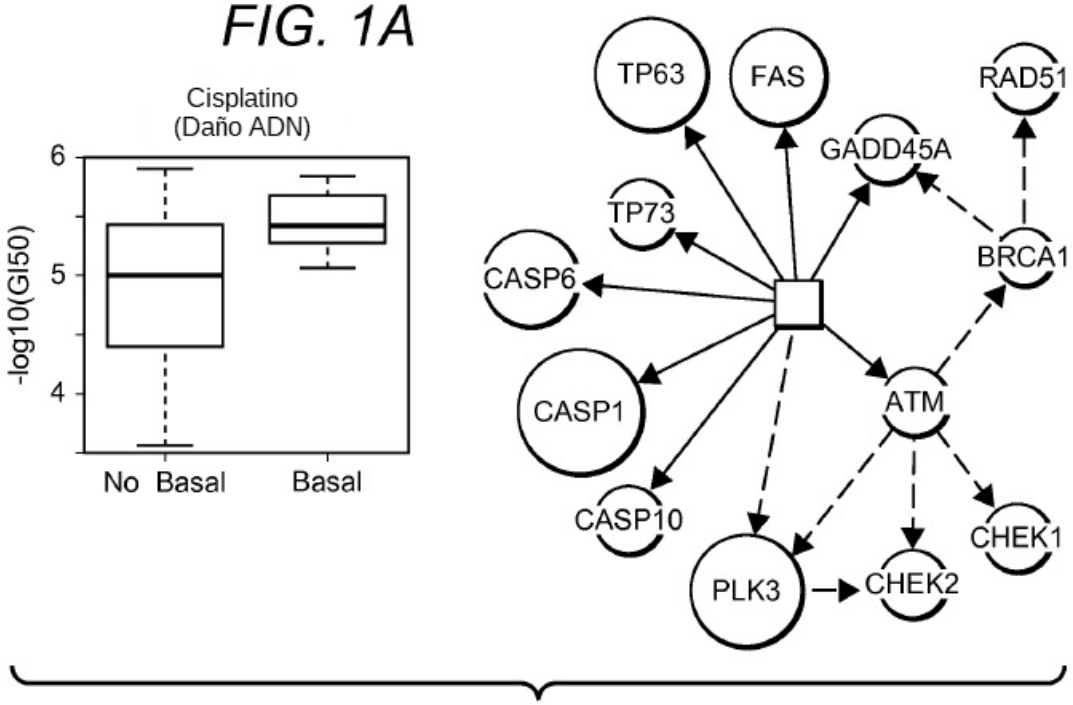
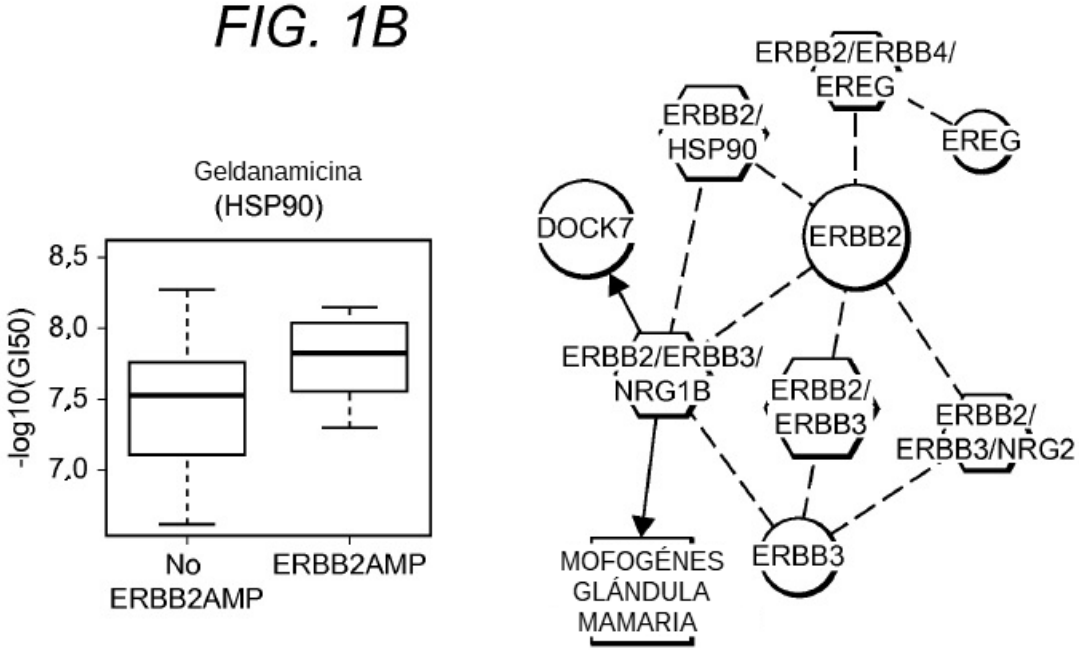

Cisplatino
(Daño ADN)
-log10(GI50)
6
5
4
No Basal
Basal
TP63
FAS
RAD51
GADD45A
TP73
BRCA1
CASP6
CASP1
ATM
CASP10
CHEK1
PLK3
CHEK2

FIG. 1B


Geldanamicina
(HSP90)
-log10(GI50)
8,5
8,0
7,5
7,0
No ERBB2AMP
ERBB2AMP
ERBB2/ERBB4/
EREG
ERBB2/
HSP90
EREG
DOCK7
ERBB2
ERBB2/ERBB3/
NRG1B
ERBB2/
ERBB3
ERBB2/
ERBB3/NRG2
MOFOGÉNES
GLÁNDULA
MAMARIA
ERBB3

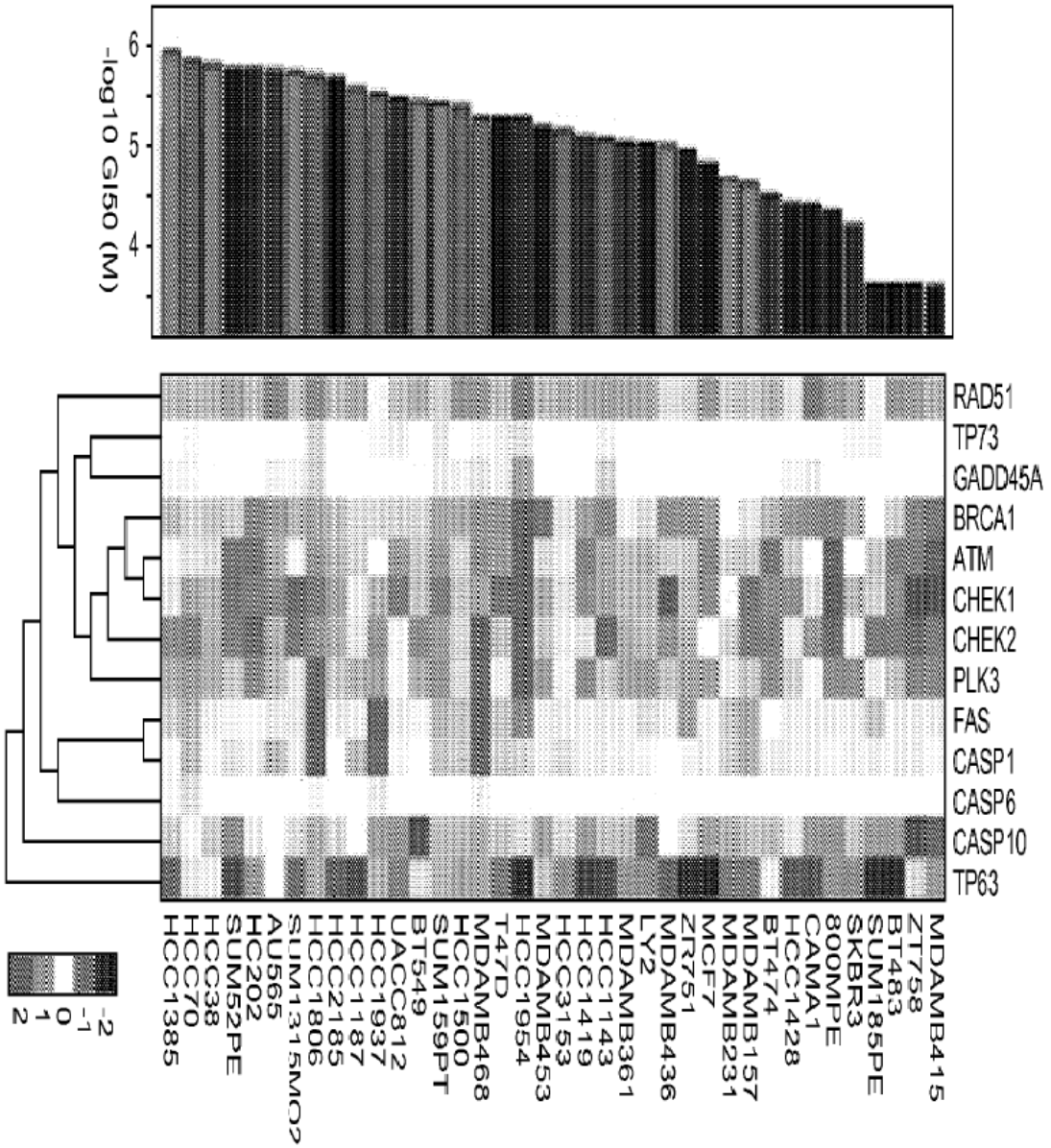

-log10 GI50 (M)
6
5
4
RAD51
TP73
GADD45A
BRCA1
ATM
CHEK1
CHEK2
PLK3
FAS
CASP1
CASP6
CASP10
TP63
HCC1385
HCC70
HCC38
SUM52PE
HC202
AU565
SUM1315MO2
HCC2185
HCC1806
HCC1187
HCC1937
UACC812
BT549
SUM159PT
HCC1500
MDAMB468
T47D
HCC1954
MDAMB453
HCC3153
HCC1419
HCC1143
MDAMB361
LY2
MDAMB436
ZR751
MCF7
MDAMB231
MDAMB157
BT474
HCC1428
CAMA1
800MPE
SKBR3
SUM185PE
BT483
ZT758
MDAMB415
2
1
0
-1
-2

FIG. 1C

FIG. 2A

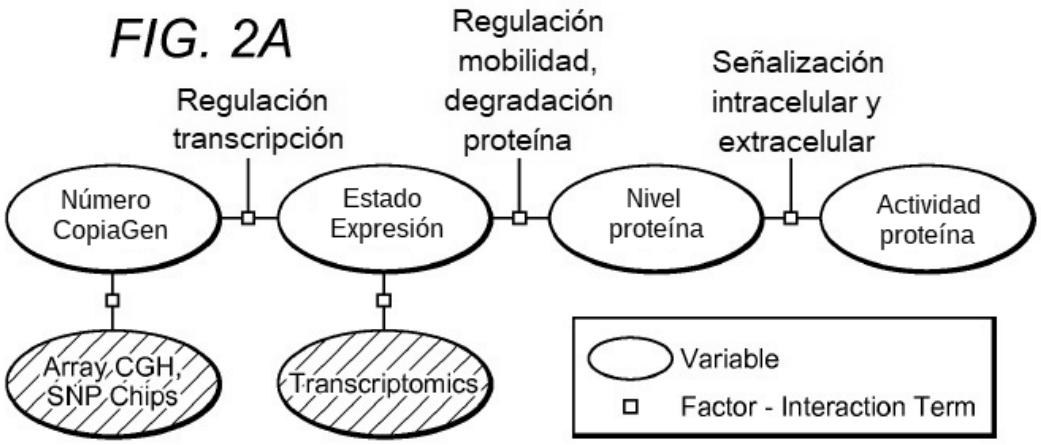


FIG. 2B

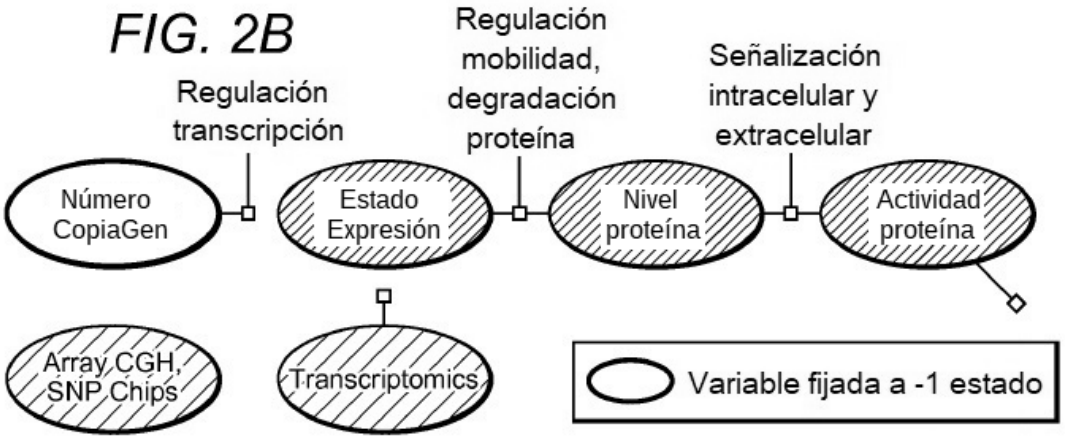
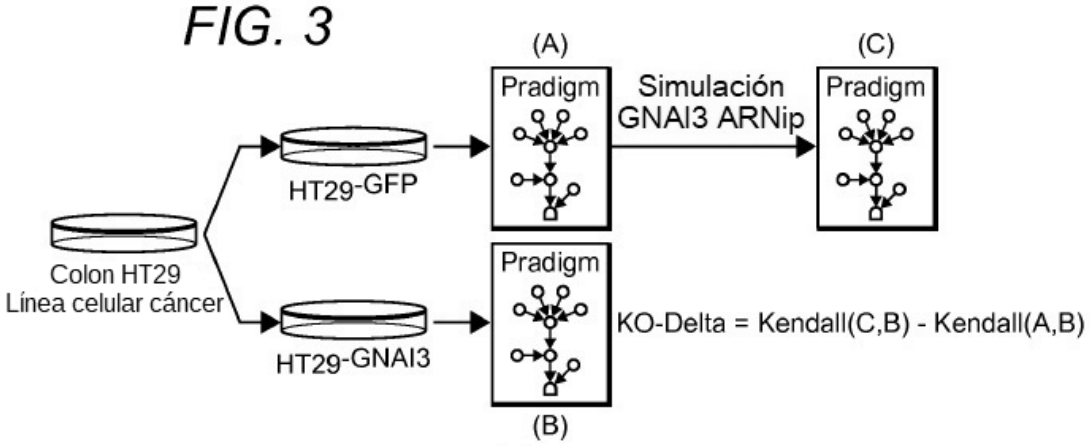
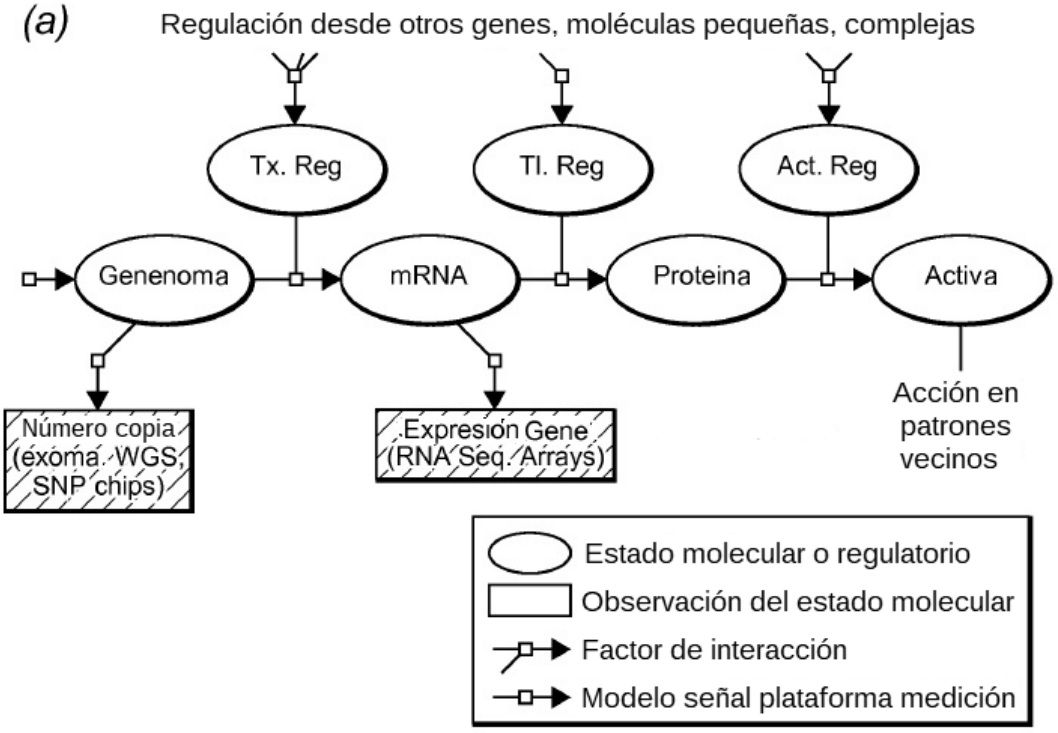


FIG. 3



Pruebas T emparejadas ({Kendall(Ci,Bj) | i ϵGFP, j ϵKO}, {Kendall(Ai,Bj) | i ϵGFP, j ϵKO})
Val-P cruzado 32 y KO pares: 0,008133


(a)
Regulación desde otros genes, moléculas pequeñas, complejas
Tx. Reg
Tl. Reg
Act. Reg
Genenoma
mRNA
Proteina
Activa
Número copia (exoma, WGS, SNP chips)
Expresión Gene (RNA Seq, Arrays)
Acción en patrones vecinos
Estado molecular o regulatorio
Observación del estado molecular
Factor de interacción
Modelo señal plataforma medición

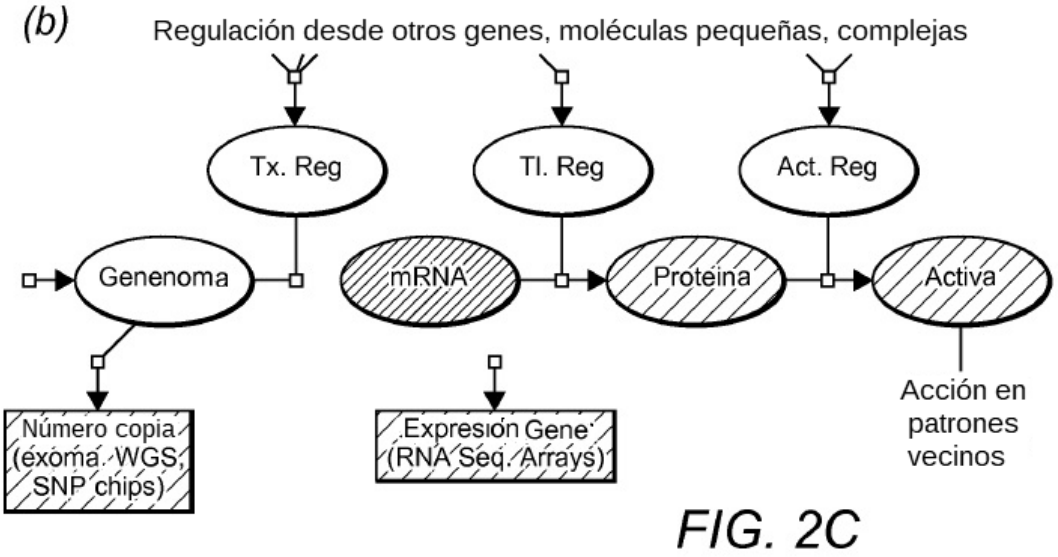

(b)
Regulación desde otros genes, moléculas pequeñas, complejas
Tx. Reg
Tl. Reg
Act. Reg
Genenoma
mRNA
Proteina
Activa
Número copia (exoma, WGS, SNP chips)
Expresión Gene (RNA Seq, Arrays)
Acción en patrones vecinos

FIG. 2C

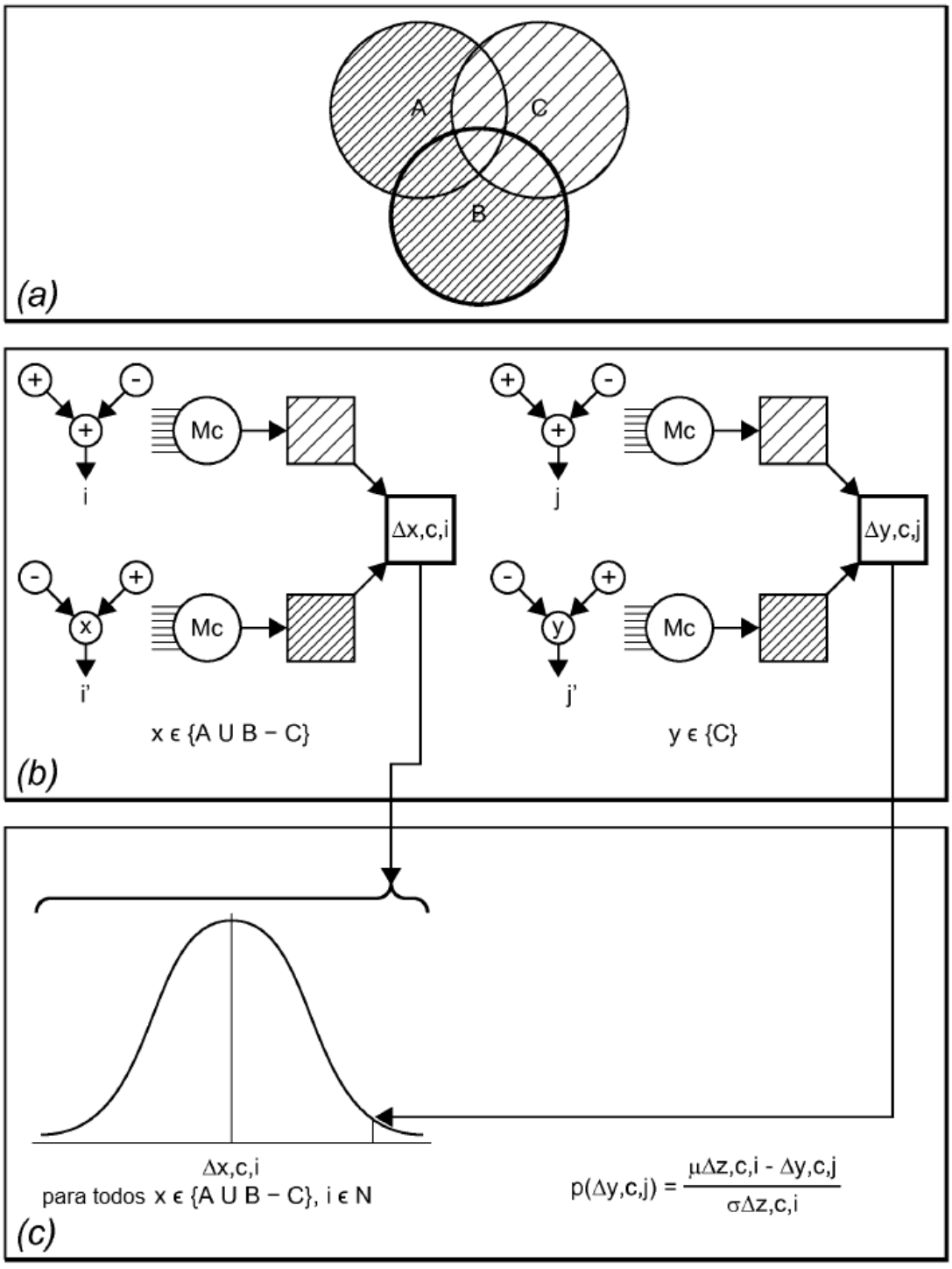
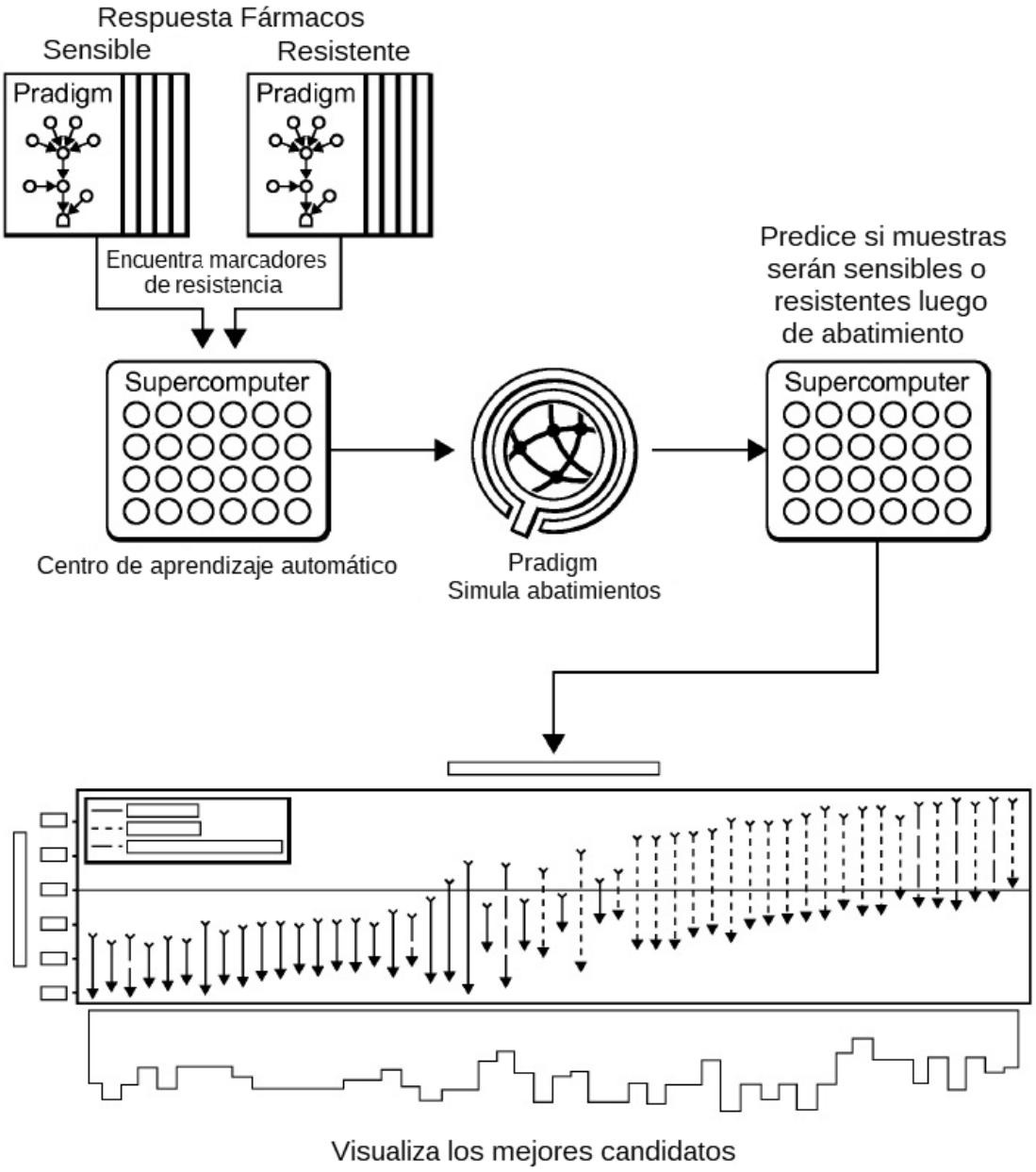

A
C
B
(a)
Mc
i
Δx,c,i
x
i'
x ϵ {A U B − C}
j
Δy,c,j
y
j'
y ϵ {C}
(b)
Δx,c,i
para todos x ϵ {A U B − C}, i ϵ N
p(Δy,c,j) = μΔz,c,i - Δy,c,j / σΔz,c,i
(c)

FIG. 2D

FIG. 4


Respuesta Fármacos
Sensible
Resistente
Pradigm
Pradigm
Encuentra marcadores de resistencia
Predice si muestras serán sensibles o resistentes luego de abatimiento
Supercomputer
Supercomputer
Centro de aprendizaje automático
Pradigm
Simula abatimientos
Visualiza los mejores candidatos

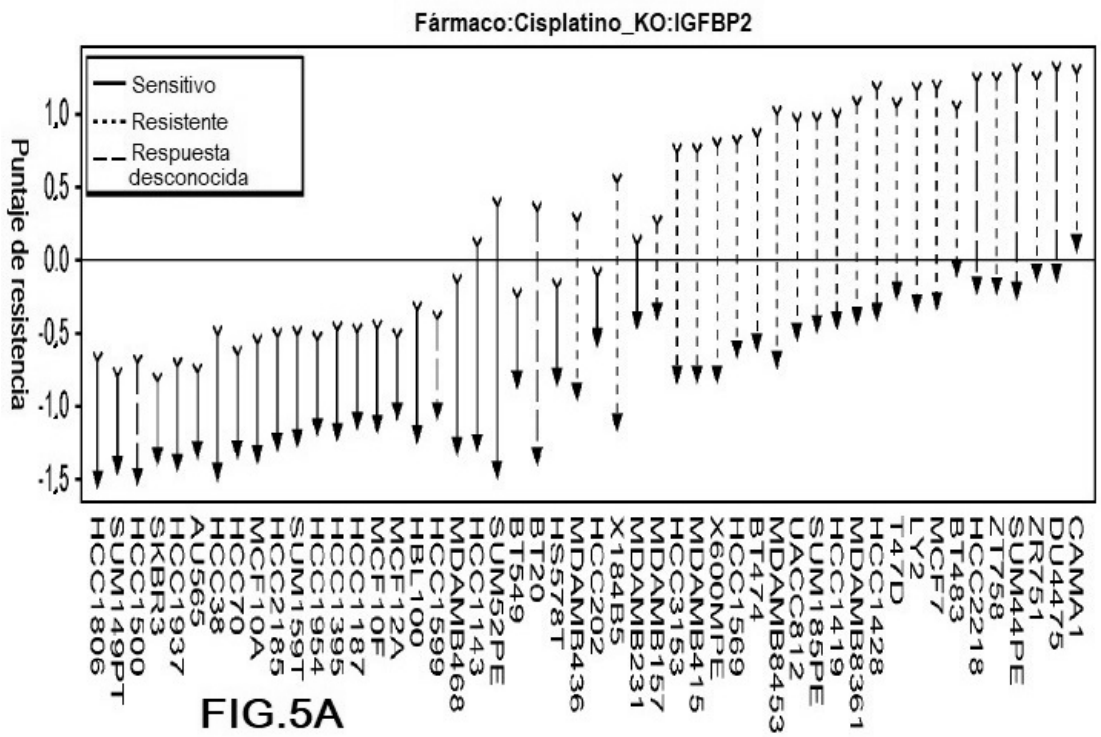

Fármaco:Cisplatino_KO:IGFBP2
Sensitivo
Resistente
Respuesta desconocida
Puntaje de resistencia
1,0
0,5
0.0
-0,5
-1,0
-1,5
HCC1806
SUM149PT
HCC1500
SKBR3
HCC1937
AU565
HCC38
HCC70
MCF10A
HCC2185
SUM159T
HCC1954
HCC1395
HCC1187
MCF10F
MCF12A
HBL100
HCC1599
HCC1143
MDAMB468
SUM52PE
BT549
BT20
HS578T
MDAMB436
HCC202
X184B5
MDAMB231
MDAMB157
HCC3153
MDAMB415
X600MPE
HCC1569
BT474
MDAMB8453
UACC812
SUM185PE
HCC1419
MDAMB8361
HCC1428
T47D
LY2
MCF7
BT483
HCC2218
ZT758
SUM44PE
ZR751
DU4475
CAMA1

FIG.5A

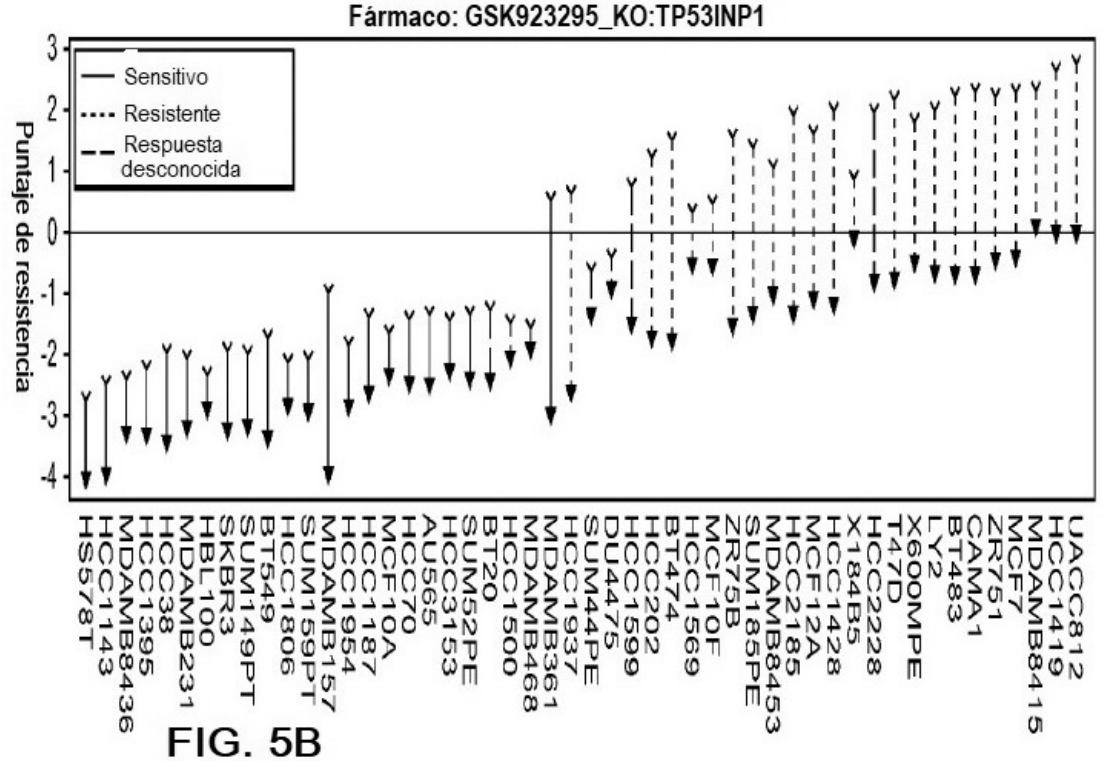

Fármaco: GSK923295_KO:TP53INP1
Sensitivo
Resistente
Respuesta desconocida
Puntaje de resistencia
3
2
1
0
-1
-2
-3
-4
HS578T
HCC1143
MDAMB436
HCC38
HCC1395
MDAMB231
HBL100
SKBR3
SUM149PT
BT549
HCC1806
SUM159PT
MDAMB157
HCC1954
HCC1187
MCF10A
HCC70
AU565
HCC3153
SUM52PE
BT20
HCC1500
MDAMB468
MDAMB361
HCC1937
SUM44PE
DU4475
HCC1599
HCC202
BT474
HCC1569
MCF10F
ZR75B
SUM185PE
MDAMB8453
HCC2185
MCF12A
HCC1428
X184B5
HCC2228
T47D
X600MPE
LY2
BT483
CAMA1
ZR751
MCF7
MDAMB8415
HCC1419
UACC812

FIG. 5B

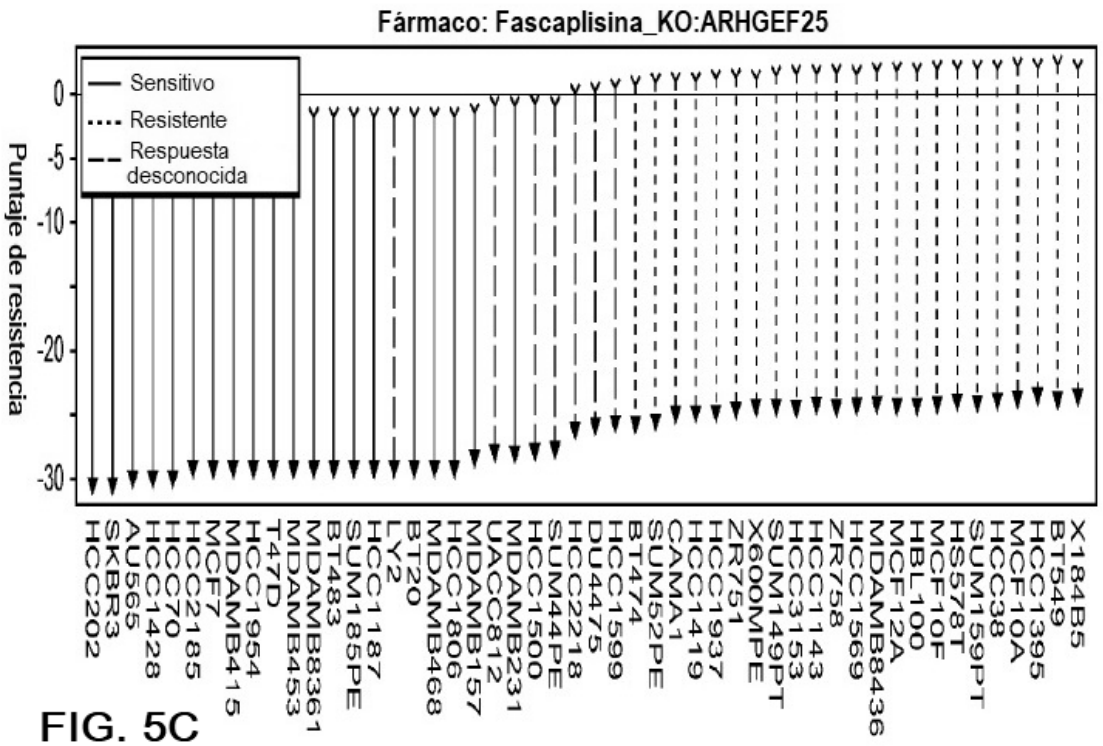

Fármaco: Fascaplisina_KO:ARHGEF25
Sensitivo
Resistente
Respuesta desconocida
Puntaje de resistencia
0
-5
-10
-20
-30
HCC202
SKBR3
AU565
HCC1428
HCC70
HCC2185
MCF7
MDAMB415
HCC1954
T47D
MDAMB453
MDAMB8361
BT483
SUM185PE
HCC1187
LY2
BT20
MDAMB468
HCC1806
MDAMB157
UACC812
MDAMB231
HCC1500
SUM44PE
HCC2218
DU4475
HCC1599
BT474
SUM52PE
CAMA1
HCC1419
HCC1937
ZR751
X600MPE
SUM149PT
HCC3153
HCC1143
ZR758
HCC1569
MDAMB8436
MCF12A
HBL100
MCF10F
HS578T
SUM159PT
HCC38
MCF10A
HCC1395
BT549
X184B5

FIG. 5C