(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 709 053**

(51) Int. CI.:

G06F 19/12 (2011.01)
G16H 50/50 (2008.01)
G06F 19/24 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **09.10.2013 PCT/US2013/064160**

(87) Fecha y número de publicación internacional: **17.04.2014 WO14059036**

(96) Fecha de presentación y número de la solicitud europea: **09.10.2013 E 13846109 (0)**

(97) Fecha y número de publicación de la concesión europea: **19.12.2018 EP 2907039**

(54) Título: **Sistemas y métodos para aprendizaje e identificación de interacciones reguladoras en rutas biológicas**

(30) Prioridad:

09.10.2012 US 201261711491 P
26.11.2012 US 201261729958 P
18.01.2013 US 201361754175 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
15.04.2019

(73) Titular/es:

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

(72) Inventor/es:

VASKE, CHARLES JOSEPH;
SEDGEWICK, ANDREW J. y
BENZ, STEPHEN CHARLES

(74) Agente/Representante:

SÁEZ MAESO, Ana

DESCRIPCIÓN

Sistemas y métodos para aprendizaje e identificación de interacciones reguladoras en rutas biológicas

Campo de la invención

El campo de la invención es el análisis computacional de datos ómicos, y particularmente en lo que se refiere a algoritmos de aprendizaje y uso del análisis de rutas.

Antecedentes de la invención

Con el advenimiento del cribado genómico de alto rendimiento, se obtuvieron conjuntos de datos cada vez más grandes que capturan el estado molecular de las células, y estos avances permitieron una mayor identificación y comprensión de los mecanismos celulares que están alterados en el cáncer. Por ejemplo, la identificación de objetivos clave frecuentemente alterados dentro de tumores específicos llevó al desarrollo de más de 40 terapias dirigidas en los últimos 20 años. Desafortunadamente, en la mayoría de los casos, la tasa de respuesta de muchos de estos fármacos es inferior al 50%, lo que destaca la comprensión incompleta de las rutas afectadas por estos fármacos. Un ejemplo típico de un mecanismo de resistencia es la activación de la ruta RAS en tumores de cáncer de colon alterados por EGFR, en el que KRAS mutado activa constitutivamente la cascada de RAS que ofrece señales de crecimiento que son independientes de la ruta EGFR, haciendo que las terapias de bloqueo de EGFR tales como la terapia con cetuximab sean en gran medida ineficaces. Por lo tanto, parece que el conocimiento de la interferencia de la ruta con cetuximab es incompleto con respecto a las rutas clave a través de las cuales las señales oncogénicas viajan dentro de las redes de señalización celular.

Tal conocimiento incompleto aparente es aún más desconcertante, ya que numerosas herramientas computacionales para integrar datos ómicos a nivel de la ruta ahora están disponibles. Entre varias otras herramientas, varios algoritmos (por ejemplo, GSEA, SPIA y PathOlogist) son capaces de identificar con éxito las rutas de interés alteradas utilizando rutas seleccionadas a partir de la literatura. Aún 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 reciben información transcripcional del gen (y el número de copia, en el caso de CONEXIC) para identificar así los posibles conductores transcripcionales a través de un conjunto de muestras de cáncer. Sin embargo, estas herramientas no intentan agrupar diferentes conductores en redes funcionales que identifican objetivos singulares de interés. Algunos algoritmos de ruta más nuevos, tales como NetBox y Mutual Exclusivity Modules in Cancer (MEMo), intentan resolver el problema de la integración de datos en el cáncer para identificar así redes a través de múltiples tipos de datos que son clave para el potencial oncogénico de las muestras. Si bien estas herramientas permiten al menos cierta integración limitada a través de las rutas para encontrar una red, generalmente no proporcionan información regulatoria ni asociación de dicha información con uno o más efectos en las rutas relevantes o en la red de rutas. Del mismo modo, GIENA busca interacciones genéticas desreguladas dentro de una ruta biológica única, pero no tiene en cuenta la topología de la ruta o el conocimiento previo sobre la dirección o la naturaleza de las interacciones.

En el análisis genómico externo, los modelos gráficos probabilísticos se han utilizado ampliamente en el análisis de redes con usos emblemáticos en forma de redes bayesianas y campos aleatorios de Markov. Varios métodos han aprendido con éxito las interacciones de los datos a través de muchos medios diferentes, incluidas las redes de relevancia. Más recientemente, PARADIGM (algoritmo de reconocimiento de ruta que usa la integración de datos en modelos genómicos) es una herramienta de análisis genómico descrita en los documentos WO2011/139345 y WO/2013/062505 y utiliza un modelo gráfico probabilístico para integrar múltiples tipos de datos genómicos en bases de datos de rutas seleccionadas. Este sistema modelo permite ventajosamente que las muestras individuales sean evaluadas solas o en el contexto de una cohorte de interés. Sin embargo, el aprendizaje de parámetros de expectativa-maximización (EM) en esa herramienta solo se realizó de forma predeterminada en los parámetros de datos de observación, ya que el tamaño limitado de los conjuntos de datos disponibles impidió una estimación robusta de los parámetros de interacción. En consecuencia, la herramienta no permitió un análisis de la interacción y la interrelación de múltiples factores que influirían en la actividad en un segmento de ruta particular, y como tal no pudo proporcionar una resolución mejorada del flujo de señal a través de redes de señalización celular.

Por lo tanto, aunque se conocen en la técnica numerosos sistemas y métodos de aprendizaje e identificación de interacciones reguladoras en rutas biológicas, todos o casi todos tienen una o más desventajas. Por ejemplo, hasta ahora las herramientas analíticas conocidas no identifican la fuerza y la dirección de las interacciones de los parámetros que modulan la actividad en una trayectoria de una ruta, y con eso no solo no permiten la predicción del flujo de señal y/o la interferencia de las actividades de la ruta, sino que también fallan para identificar el posible uso diferencial de los parámetros o elementos de la ruta. Desde una perspectiva diferente, las herramientas actualmente conocidas normalmente solo consideran actividades de genes individuales, pero no examinan las estadísticas relacionadas con los enlaces reguladores y, por lo tanto, solo proporcionan un modelo estático en lugar de un modelo dinámico. En consecuencia, los modelos conocidos tampoco permitirán examinar cómo los diferentes reguladores dentro de una red pueden producir fenotipos celulares similares a pesar de usar rutas completamente diferentes para lograrlos. Por lo tanto, subsiste la necesidad de sistemas y métodos mejorados para el aprendizaje e identificación de interacciones reguladoras en rutas biológicas.

Sumario de la invención

La presente invención se refiere a un método implementado por ordenador para clasificar un tejido como perteneciente a un tejido específico de subtipo, que comprende:

(a) obtener, a través de un módulo de interfaz de entrada ómico (120), al menos un conjunto de datos ómico (135) representativo del tejido;

(b) acceder, a través de un módulo de procesamiento ómico (170), a un modelo de ruta biológica (150) que tiene una pluralidad de elementos de la ruta que comprenden al menos uno de una secuencia de ADN, una secuencia de ARN, una proteína y una función de proteína, en la que al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta en función de una pluralidad de parámetros reguladores;

i) cuando el elemento de la ruta comprende una secuencia de ADN, al menos uno de la pluralidad de parámetros reguladores se selecciona del grupo que consiste en un factor de transcripción, un activador de la transcripción, una subunidad de ARN polimerasa, un elemento regulador en cis, un elemento regulador en trans, una histona acetilada, una histona metilada y un represor,

ii) cuando el elemento de la ruta comprende una secuencia de ARN, al menos uno de la pluralidad de parámetros reguladores se selecciona del grupo que consiste en un factor de iniciación, un factor de traducción, una proteína de unión a ARN, una proteína ribosómica, un ARNpi y una proteína de unión a poliA, y

iii) cuando el elemento de la ruta comprende una proteína, al menos uno de la pluralidad de parámetros reguladores es una fosforilación, una acilación, una escisión proteolítica y asociación con al menos una segunda proteína;

(c) inferir, mediante el módulo (170) de procesamiento ómico, basado en al menos un conjunto de datos (135) ómico y el modelo (150) de ruta, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores, en el que las probabilidades condicionales de enlaces individuales se aprenden y se utiliza una suposición de Bayes sencilla para calcular la probabilidad de un nodo hijo Y dados los progenitores X_1, ... X_n, en donde la probabilidad F se calcula con base en la expresión:

$$F(Y|X_1, \ldots, X_N) = \frac{1}{Z} P(Y) \prod_i P(X_i|Y)$$

en la que Z es una constante de normalización que corresponde a $P(X_1,...X_n)$, siendo el modelo de ruta un modelo probabilístico configurado para usar gráficos de factores usando un modelo de regulación independiente;

(d) actualizar el modelo de ruta (150) basado en las correlaciones de interacción aprendidas, en el que la interfaz de entrada ómica es una interfaz de computación configurada para recibir uno o más conjuntos de datos ómicos, y en el que el módulo de procesamiento ómico es una parte de un dispositivo de computación, en el que una prueba G determina la significación estadística de la dependencia entre los progenitores que proporcionan una distribución de hijos, y

en el que una correlación de Pearson o información mutua puntual ponderada (WPMI) determina el signo de interacción para los parámetros reguladores,

(e) hacer coincidir el conjunto derivado de correlaciones de interacción con un conjunto conocido a priori de correlaciones de interacción que está asociado con un tejido específico de subtipo conocido; y

(f) utilizar el emparejamiento para clasificar que el conjunto de datos ómico representativo del tejido pertenece al tejido específico de subtipo conocido, en el que el subtipo incluye tejido resistente al fármaco, tejido metastático, tejido tratado con fármaco o una variante clonal de un tejido.

Los conjuntos de datos ómicos pueden comprender datos (135) del genoma completo, datos del genoma parcial u objetos de secuencia diferencial, y en los que los conjuntos de datos (135) ómicos se obtienen de una base de datos (130) genómica, un servidor (130) BAM, o un dispositivo (130) de secuenciación.

La invención también se refiere a un aparato configurado para llevar a cabo el método de la invención.

Las correlaciones de interacción entre los parámetros reguladores se deducen con base en un conjunto de datos ómico y/o el modelo de ruta. Las correlaciones de interacción identificadas ahora permiten identificar la fuerza y la dirección de las interacciones de los parámetros que modulan la actividad en una trayectoria de una ruta. En consecuencia, los sistemas y métodos contemplados permiten la predicción del flujo de señal y/o la interferencia de las actividades de la ruta, así como el uso potencialmente diferencial de los parámetros o elementos de la ruta. Visto desde una perspectiva diferente, los sistemas y métodos contemplados proporcionan un modelo de ruta dinámica que se puede utilizar para la identificación del flujo de señal (incluso diferencial) a través de una o más rutas, así como la predicción del flujo de señal en varios escenarios (reales o simulados).

En un aspecto de esta divulgación, un motor de aprendizaje comprende una interfaz de entrada ómica que recibe uno o más conjuntos de datos ómicos (por ejemplo, datos del genoma completo, datos del genoma parcial u objetos de secuencia diferencial). Un módulo de procesamiento ómico está acoplado con la interfaz y está configurado para (a) acceder a un modelo de ruta que tiene una pluralidad de elementos de ruta (por ejemplo, secuencia de ADN, secuencia de ARN, proteína, función de proteína) en el que dos o más de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores, (b) obtener, a través de la interfaz de entrada ómica, al menos uno de los conjuntos de datos ómicos, (c) inferir, basado en al menos un conjunto de datos ómico y el modelo de ruta, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores, y (d) actualizar el modelo de ruta en función de las correlaciones de interacción. Más típicamente, el motor de aprendizaje comprende además o está acoplado a una base de datos genómica, un servidor BAM o un dispositivo de secuenciación.

En algunas realizaciones, el elemento de ruta comprende una secuencia de ADN y el parámetro regulador es un factor de transcripción, un activador de la transcripción, una subunidad de ARN polimerasa, un elemento regulador en cis, un elemento regulador en trans, una histona acetilada, una histona metilada, y/o un represor. En otras realizaciones, el elemento de la ruta comprende una secuencia de ARN y el parámetro regulador es un factor de iniciación, un factor de traducción, una proteína de unión a ARN, una proteína ribosómica, un ARNpi y/o una proteína de unión a poliA, y en otras realizaciones, el elemento de la ruta comprende una proteína y el parámetro regulador es una fosforilación, una acilación, una escisión proteolítica y una asociación con al menos una segunda proteína.

En los aspectos especialmente preferidos de esta divulgación, el módulo de procesamiento de ómicos está configurado para inferir la correlación de interacción utilizando un modelo probabilístico, que utiliza un modelo de regulación codependiente y/o independiente. Además, generalmente se prefiere que el modelo probabilístico determine además una significancia de dependencia entre la pluralidad de los parámetros reguladores y la actividad de la ruta y/o una significancia de dependencia condicional entre los parámetros reguladores dada una actividad de la ruta. Adicionalmente, se contempla que el modelo probabilístico determine además el signo de interacción para los parámetros reguladores.

Por lo tanto, y vistos desde una perspectiva diferente, en esta divulgación los inventores también contemplan un método para generar un modelo de ruta que incluye una etapa para obtener, a través de una interfaz de entrada ómica, al menos un conjunto de datos ómico (por ejemplo, datos del genoma completo, datos del genoma parcial, u objetos de secuencia diferencial). Los métodos contemplados también incluyen otra etapa para acceder, a través de un módulo de procesamiento ómico, a un modelo de ruta que tiene una pluralidad de elementos de ruta en los que al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores, y una etapa adicional para inferir, a través del módulo de procesamiento ómico, basado en al menos un conjunto de datos ómico y el modelo de ruta, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores. En otra etapa más, el modelo de ruta se actualiza en función de las correlaciones de interacción. Normalmente, los conjuntos de datos ómicos se obtienen de una base de datos genómica, un servidor BAM o un dispositivo de secuenciación.

En aspectos adicionales de esta divulgación, la etapa de inferir se basa en un modelo probabilístico, y más preferiblemente el modelo probabilístico usa un modelo de regulación codependiente y/o independiente. Además, los métodos contemplados incluyen una etapa para determinar el significado de la dependencia entre la pluralidad de los parámetros reguladores y la actividad de la ruta y/o el significado de la dependencia condicional entre los parámetros reguladores dada una actividad de la ruta. También se prefiere además que se contemple que tales métodos incluyen una etapa de determinación del signo de interacción para los parámetros reguladores.

En otros aspectos de esta divulgación, un método para identificar correlaciones de interacción específicas de subtipo para parámetros reguladores de un nodo regulador en un modelo de ruta incluye una etapa para obtener, a través de una interfaz de entrada ómica, al menos un conjunto de datos ómico representativo de un tejido de subtipo y una etapa adicional para acceder, a través de un módulo de procesamiento ómico, al modelo de ruta que tiene una pluralidad de elementos de ruta en los cuales al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene el nodo regulador que controla la actividad a lo largo de la trayectoria en función de la pluralidad de parámetros reguladores. Los métodos contemplados incluyen además una etapa para derivar las correlaciones de interacción de subtipo, a través del módulo de procesamiento ómico, de al menos un conjunto de datos ómico representativo del tejido de subtipo mediante análisis de probabilidad de interacciones entre la pluralidad de parámetros reguladores, y otra etapa de presentar las correlaciones de interacción derivadas de subtipo en el modelo de ruta. En aspectos especialmente preferidos, el tejido de subtipo es un tejido resistente al fármaco, un tejido metastásico, un tejido tratado con fármaco o una variante clonal de un tejido.

Cuando se desee, los métodos contemplados pueden incluir además una etapa de validación de las correlaciones de interacción derivadas de subtipos utilizando al menos un experimento *in vitro, in silico* e *in vivo*.

En otros aspectos adicionales de la materia de la invención, los inventores contemplan un método para clasificar un conjunto de datos ómico representativo de un tejido como el perteneciente a un tejido específico de subtipo. Tales métodos típicamente comprenderán una etapa para obtener, a través de una interfaz de entrada ómica, el conjunto de datos ómico representativo del tejido, y otra etapa para derivar, para el conjunto de datos ómico, un conjunto de

correlaciones de interacción entre una pluralidad de parámetros reguladores de un nodo regulador en un modelo de ruta. En otra etapa más, el conjunto derivado de correlaciones de interacción se hace coincidir con un conjunto conocido a priori de correlaciones de interacción que se asocia con un tejido específico de subtipo conocido, y luego se usa la coincidencia para clasificar que el conjunto de datos ómico representativo del tejido pertenece al tejido específico de subtipo conocido.

Lo más preferiblemente, la etapa de obtención comprende generar el conjunto de datos ómico representativo del tejido a partir de una muestra de tejido (por ejemplo, una muestra de tumor) de un tejido con una característica reguladora desconocida, y el tejido específico de subtipo conocido es un tejido resistente al fármaco, un tejido metastásico, un tejido tratado con fármacos o una variante clonal de un tejido.

En otro aspecto más de esta divulgación, los inventores contemplan un método para identificar un objetivo que se puede tratar con fármacos en un modelo de ruta que tiene una pluralidad de elementos de ruta en los que al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores. Tales métodos incluirán las etapas de (a) obtener, a través de una interfaz de entrada ómica, un conjunto de datos ómico representativo de un tejido, (b) derivar, para el conjunto de datos ómico, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores del nodo regulador en el modelo de ruta, y (c) identificar un fármaco que afecta la actividad de la ruta en la que se predice que el fármaco interfiere con las correlaciones de interacción. Más típicamente, el nodo regulador afecta al menos a una de una modificación de transcripción, traducción y postraduccional de una proteína, y el fármaco es un fármaco disponible comercialmente y tiene un modo de acción conocido.

En otro aspecto más de esta divulgación, los inventores contemplan un método para identificar una ruta objetivo en un modelo de ruta que tiene una pluralidad de elementos de ruta en los que al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores. Dichos métodos comprenderán preferiblemente una etapa para obtener, a través de una interfaz de entrada ómica, un conjunto de datos ómico representativo de un tejido, una etapa adicional para derivar, para el conjunto de datos ómico, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores del nodo regulador en el modelo de ruta, y una etapa adicional para identificar una ruta como la ruta objetivo basada en un efecto conocido de un fármaco en la correlación de interacción.

Lo más preferiblemente, el efecto conocido es al menos uno de un efecto inhibidor sobre una quinasa, un efecto inhibidor sobre un receptor y un efecto inhibidor sobre la transcripción. Entre otras rutas objetivo adecuadas, las rutas objetivo especialmente contempladas incluyen una ruta regulada de calcio/calmodulina, una ruta de citoquina, una ruta de quimioquina, una ruta regulada del factor de crecimiento, una ruta regulada de hormonas, una ruta regulada de MAP quinasa, una ruta regulada de fosfatasa y una ruta regulada de Ras. Dichos métodos pueden incluir además una etapa para proporcionar un consejo de tratamiento basado en la ruta identificada.

Por lo tanto, los métodos contemplados también incluirán un método para simular in silico un efecto de tratamiento de un fármaco que incluye una etapa para obtener un modelo de ruta que tiene una pluralidad de elementos de ruta en los que al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores. Los métodos contemplados incluirán además una etapa para identificar un fármaco que se sabe que afecta al menos a un parámetro regulador, y otra etapa para alterar in silico, a través de un módulo de procesamiento ómico y basado en el efecto conocido del fármaco, al menos uno del nodo regulador, la actividad, y al menos los parámetros reguladores en el modelo de ruta, y otra etapa más para determinar un efecto secundario de la alteración en el modelo de ruta. En general, el efecto secundario se encuentra en otro nodo regulador, otra actividad y otro parámetro regulador en el modelo de ruta.

Varios objetivos, características, aspectos y ventajas del objeto 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 adjuntos en las que números similares representan componentes similares.

Breve divulgación del dibujo

La Figura 1 es un ejemplo de una ilustración esquemática de un motor de aprendizaje de acuerdo con el objeto de la invención.

La Figura 2A es un ejemplo de una ilustración esquemática de una estructura gráfica de factores de acuerdo con el objeto de la invención, y la Figura 2B muestra esquemáticamente modelos de regulación alternativos para los nodos de transcripción, traducción y activación.

La Figura 3A es un ejemplo de un gráfico del análisis del componente principal (PCA) de los vectores WPMI para cada interacción aprendida en toda la cohorte de TCGA. La Figura 3B ilustra la membresía del grupo de enlaces significativos marcados como activación e inhibición en la ruta, y la Figura 3C muestra mapas de calor de los valores de WPMI de los centroides de los grupos que muestra un rango de inhibición fuerte a la activación fuerte.

Las Figuras 4A y 4B son diagramas de barras de membresías del grupo para los valores de WPMI de enlaces

significativos en la inicialización informativa (4A) y llana (4B).

La Figura 5A es un ejemplo de un gráfico que muestra el porcentaje de nodos hijos únicos que fallan en las siguientes pruebas en cada etapa EM de un proceso que aprende una probabilidad condicional completa, y la Figura 5B es una ilustración esquemática de ejemplos de tripletes coherentes frente a incoherentes.

Las Figuras 6A-6C son ejemplos de gráficos que muestran curvas de supervivencia de Kaplan-Meier para el análisis de la ruta utilizando diferentes métodos analíticos.

La Figura 7 es un ejemplo de una representación de mapa de calor de los rangos de puntuación G.

Las Figuras 8A-8B son ejemplos de diagramas de caja que representan señales de WPMI agrupadas por tejido para los enlaces de activación de PPARA-RXRA y TAp73a.

Descripción detallada

Los inventores han descubierto ahora que se puede implementar un modelo de ruta gráfico probabilístico en el que se determina estadísticamente una interrelación de parámetros reguladores. En consecuencia, el análisis y las simulaciones de los sistemas y métodos contemplados proporcionarán una precisión significativamente mejorada, y permitirá la identificación del uso diferencial de elementos reguladores dentro de diferentes rutas y/o tejidos secundarios.

Por lo tanto, se debe tener en cuenta que al identificar enlaces reguladores con distribuciones de uso significativamente diferentes dentro de un fenotipo de interés en una cohorte, ahora es posible examinar cómo los diferentes reguladores dentro de una red podrían producir fenotipos celulares similares a pesar de utilizar rutas completamente diferentes para lograrlo. Además, los parámetros así aprendidos pueden usarse como base para pruebas estadísticas para establecer qué tan bien las muestras individuales o los subconjuntos de la cohorte siguen la distribución de los patrones de parámetros previamente aprendidos para cada nodo regulador.

A lo largo de la siguiente discusión, se harán numerosas referencias con respecto a servidores, servicios, interfaces, portales, plataformas u otros sistemas formados a partir de dispositivos informáticos. Debe apreciarse que el uso de dichos términos pretende representar uno o más dispositivos informáticos que tienen al menos un procesador configurado para ejecutar instrucciones de software almacenadas en un medio legible tangible y no transitorio por un ordenador. Por ejemplo, un servidor puede incluir uno o más ordenadores que funcionan como un servidor web, un servidor de base de datos u otro tipo de servidor de una manera que cumpla con las funciones, responsabilidades o funciones descritas.

Por ejemplo, la Fig. 1 muestra a modo de ejemplo el ecosistema 100 que incluye el motor 110 de aprendizaje. El motor 110 de aprendizaje está configurado para procesar uno o más del conjunto de datos 135 ómico en vista de uno o más del modelo 150 de ruta. El motor 110 de aprendizaje comprende dos componentes: la interfaz 120 ómica a través de la cual el motor 110 de aprendizaje obtiene los conjuntos de datos de interés y el módulo 170 de procesamiento ómico configurado para analizar los conjuntos de datos. En el ejemplo mostrado, el motor 110 de aprendizaje se ilustra como un dispositivo informático accesible a través de la red 115 (por ejemplo, Internet, WAN, LAN, VPN, National Lamba Rail (véase URL www.nlr.net), etc.), posiblemente como una granja de servidores HTTP. En algunos ejemplos, el motor 110 de aprendizaje ofrece sus servicios a través de la red 115 por una tarifa. Por ejemplo, el motor 110 de aprendizaje puede exponer una o más de las interfaces 120 de entrada ómica al analista 170 u otro usuario a través de una Plataforma como Servicio (PaaS) basada en la nube, Infraestructura como Servicio (IaaS), Software como servicio (SaaS), u otro tipo de servicio. En otras realizaciones, el motor 110 de aprendizaje podría ser un dispositivo informático local en relación con el analista 170 y estar configurado para ejecutar uno o más paquetes de instrucciones de software que cumplan los roles y responsabilidades del motor 110 de aprendizaje como se explica a continuación.

La interfaz 120 de entrada ómica representa una interfaz informática configurada para recibir uno o más conjuntos de datos 135 ómicos. Un ejemplo de la interfaz 120 podría incluir un servidor HTTP capaz de recibir conjuntos de datos 135 a través de la red 115. Por ejemplo, el conjunto de datos 135 podría incluir un archivo en un formato serializado (por ejemplo, XML), formato BAMBAM u otros formatos digitales adecuados que pueden transmitirse a través del servidor HTTP. En otras realizaciones, la interfaz 120 podría tomar la forma de una Interfaz de Programa de Aplicación (API) a través de la cual las estructuras de datos o sus referencias pueden pasarse al motor 110 de aprendizaje a través de la red 115 como una llamada de procedimiento remoto o incluso a través de una llamada de función de biblioteca local. Se debe tener en cuenta que la interfaz 120 de entrada ómica se puede configurar para acoplarse con uno o más de la fuente 130 del conjunto de datos ómico, posiblemente operando como una base de datos. En algunas realizaciones, el motor 110 de aprendizaje comprende una base de datos genómica o un dispositivo de secuenciación acoplado a la interfaz 120 de entrada ómica.

El conjunto de datos 135 ómico puede incluir un amplio espectro de datos ómicos. En realizaciones más preferidas, el conjunto de datos 135 ómico representa datos genómicos, posiblemente datos del genoma completo, datos parciales del genoma, objetos de secuencia diferencial u otros datos genómicos. Además, el conjunto de datos 135 ómico también puede representar otros tipos de datos que incluyen proteómica, metabolómica, lipidómica, cinómica u otras modalidades de datos ómicos.

El módulo 170 de procesamiento representa al menos una parte de un dispositivo de computación junto con la interfaz 120 de entrada ómica y está configurado para analizar el conjunto de datos 135 con respecto al modelo 150 de ruta. Un aspecto del módulo 170 de procesamiento incluye la capacidad de acceder a uno o más del modelo 150 de ruta, posiblemente de la base de datos 140 del modelo de ruta u otra fuente de modelo. En algunas realizaciones, el módulo 170 de procesamiento ómico también podría aprovechar la interfaz 120 de entrada ómica para acceder a la base de datos 140 del modelo de ruta.

El modelo 150 de ruta representa un modelo digital de actividad del sistema ómico objetivo a modelar, posiblemente en forma de un gráfico de factores. Cada modelo 150 de ruta comprende una pluralidad de elementos 151A a 151N de ruta, denominados colectivamente como elementos 151 de ruta. Los elementos 151 de ruta representan etapas a lo largo de una ruta en la que tiene lugar la actividad. Entre al menos dos elementos 151 de la ruta, los elementos 151A y 151B de la ruta como se muestra, por ejemplo, está un nodo regulador representado por el nodo 153A regulador, genéricamente denominado nodo 153 regulador. Aunque no se ilustra, puede haber nodos 153 reguladores adicionales entre cada conjunto de los elementos 151 de la ruta. Por lo tanto, al menos dos de los elementos 151 de la ruta, por ejemplo, los elementos 151A y 151B de la ruta, están acoplados entre sí a través de una ruta que tiene un nodo 153 regulador, el nodo 153A regulador como se muestra. El nodo 153 regulador del modelo 150 de ruta controla la actividad a lo largo de la ruta entre los elementos en función de uno o más parámetros 155A reguladores, genéricamente denominados como parámetros 155 reguladores. Se debería apreciar que el modelo 150 de ruta puede incluir cualquier número práctico de elementos 151 de ruta, los nodos 153 reguladores y los parámetros 155 reguladores. Como ejemplo, considérese los escenarios donde los elementos 151 de la ruta incluyen una secuencia de ADN, una secuencia de ARN, una proteína, una función de proteína u otros elementos de actividad.

En los escenarios en los que uno de los elementos 151 de la ruta comprende una secuencia de ADN, los parámetros 155 reguladores pueden incluir un factor de transcripción, un activador de la transcripción, una subunidad de ARN polimerasa, un elemento regulador en cis, un elemento regulador en trans, una histona acetilada, una histona metilada, un represor u otros parámetros de actividad. Además, en los escenarios en los que uno de los elementos 151 de la ruta comprende una secuencia de ARN, los parámetros 155 reguladores pueden incluir un factor de iniciación, un factor de traducción, una proteína de unión a ARN, una proteína ribosomal, un ARNpi, una proteína de unión a poliA u otro parámetro de actividad de ARN. Aún más, en escenarios en los que uno de los elementos 151 de la ruta comprende una proteína, los parámetros 155 reguladores podrían incluir la fosforilación, una acilación, una escisión proteolítica o una asociación con al menos una segunda proteína.

El módulo 170 de procesamiento ómico aprovecha el modelo 150 de ruta junto con el conjunto de datos 135 para inferir un conjunto de correlaciones 160 de interacción entre la pluralidad de parámetros reguladores. Un modelo tipo de ejemplo que puede aprovecharse para inferir correlaciones 160 de interacción incluye un modelo probabilístico en el que el modelo configura el modelo 170 de procesamiento ómico para comparar pares de parámetros reguladores en múltiples conjuntos de datos 135 sin procesar. En algunos ejemplos, los nodos 153 reguladores operan con base en un modelo de regulación dependiente donde el motor 110 de aprendizaje aprende una tabla de probabilidad condicional completa del hijo dados los progenitores. En otros casos, los nodos 153 reguladores pueden operar con base en un modelo de regulación independiente en el que el motor 110 de aprendizaje aprende las probabilidades condicionales utilizando una suposición de Bayes sencillo para calcular la probabilidad del nodo hijo dado el progenitor.

Los modelos probabilísticos contemplados se configuran además para determinar el significado de la dependencia entre la pluralidad de parámetros 155 reguladores y la actividad de la ruta correspondiente, o la significación de la dependencia condicional entre los parámetros reguladores dada una actividad de la ruta. Por ejemplo, una vez que se calculan o establecen las probabilidades condicionales, el módulo 150 de procesamiento ómico puede utilizar una prueba G para determinar el significado. Además, el modelo probabilístico se puede configurar además para determinar el signo de interacción de los parámetros reguladores. Una vez que se establecen las correlaciones 160 de interacción, el modelo 150 de ruta se puede actualizar para reflejar las relaciones de interacción aprendidas. En consecuencia, debe apreciarse que un motor de aprendizaje normalmente comprenderá una interfaz de entrada ómica que recibe uno o más conjuntos de datos ómicos. Dicha interfaz de entrada ómica se puede acoplar a una variedad de dispositivos o sistemas que, en la mayoría de los casos típicos, proporcionarán información ómica a un módulo de procesamiento ómico. Por ejemplo, la información ómica se puede derivar de los datos publicados, las bases de datos genómicas, RNómica y/o proteómicas, de los archivos de salida de las bases de datos de información ómica (por ejemplo, TCGA), así como de otros dispositivos, servicios y redes que proporcionan datos ómicos, incluidas las bases de datos de secuencias de ADN, ARN y/o proteínas, dispositivos de secuenciación, servidores BAM, etc. En consecuencia, debe apreciarse que el formato de los datos puede cambiar considerablemente y puede presentarse como datos del genoma completo, datos del genoma parcial u objetos de secuencia diferencial.

En la mayoría de los casos, el módulo de procesamiento ómico está acoplado informativamente con la interfaz y está configurado para (a) acceder a un modelo de ruta que tiene una pluralidad de elementos de ruta (por ejemplo, secuencia de ADN, secuencia de ARN, proteína, función de proteína) en la que dos o más de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores, (b) obtener, a través de la interfaz de entrada ómica, al menos uno de los conjuntos de datos ómicos, (c) inferir, basado en al menos un conjunto de datos ómico y el modelo de ruta, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores, y (d) actualizar el modelo de ruta en función de las correlaciones de interacción.

Se debe reconocer que los modelos de ruta para (a) pueden generarse a partir de un conjunto de datos ómicos, o pueden obtenerse a partir de determinaciones previas. Por lo tanto, los sistemas y métodos contemplados incluirán un módulo de almacenamiento que está acoplado al módulo de procesamiento ómico, en el que el módulo de almacenamiento almacena uno o más modelos de ruta determinados previamente. También debe reconocerse que los modelos de ruta almacenados pueden corresponder a tejido "normal" o a tejido enfermo. Cuando el modelo de la ruta es de un tejido enfermo, también debe apreciarse que el tejido enfermo puede ser de un subtipo particular que se caracteriza por un rasgo secundario (por ejemplo, un subtipo que es resistente al tratamiento con un fármaco particular, subtipo que proviene del tejido metastásico, etc.). También se contempla que los datos ómicos se pueden proporcionar a través de la interfaz de muchas maneras. Por ejemplo, los datos pueden proporcionarse en un solo archivo, o en una colección de archivos distintos, que pueden ser proporcionados por un proveedor de servicios, desde una biblioteca almacenada previamente, o desde un dispositivo de secuenciación o sistema de análisis de secuencias. Por lo tanto, el motor de aprendizaje puede comprender además o puede estar acoplado a una base de datos genómica, un servidor BAM o un dispositivo de secuenciación.

Según la ruta en particular, se debe tener en cuenta que la naturaleza del elemento de la ruta cambiará considerablemente, y con ello la naturaleza del parámetro regulador. En general, se debe tener en cuenta, sin embargo, que el parámetro regulador determinará el flujo de una señal a través de la ruta desde el elemento de la ruta a un elemento más adelante. Por ejemplo, cuando el elemento de la ruta es o comprende una secuencia de ADN, los parámetros reguladores contemplados serán aquellas entidades celulares que afectan la transcripción (u otra función) de la secuencia de ADN. Por lo tanto, los parámetros reguladores contemplados para una secuencia de ADN incluyen uno o más factores de transcripción, activadores de transcripción, subunidades de ARN polimerasa, elementos reguladores en cis, elementos reguladores en trans, histonas (des)acetiladas, histonas (des)metiladas y/o represores. Del mismo modo, cuando el elemento de la ruta es o comprende una secuencia de ARN, se contempla que los parámetros reguladores adecuados incluyen factores que afectan la traducción (u otra actividad) del ARN. En consecuencia, tales parámetros reguladores incluyen factores de iniciación, factores de traducción, proteínas de unión a ARN, ARN ribosomal y/o proteínas, ARNpi y/o proteínas de unión a poliA. De la misma manera, en el presente documento el elemento de la ruta es o comprende una proteína, todos los factores que afectan la actividad de esa proteína se consideran parámetros reguladores adecuados y, por lo tanto, pueden incluir otras proteínas (por ejemplo, que interactúan con la proteína para formar un complejo activado o complejo con actividad diferencial), modificación química (por ejemplo, fosforilación, acilación, escisión proteolítica, etc.).

Con respecto a la inferencia del conjunto de correlaciones de interacción entre los parámetros reguladores, generalmente se contempla que dicha inferencia se basa en el conjunto de datos ómico y/o el modelo de ruta, y también se contempla generalmente en esta divulgación que la inferencia se realiza utilizando un modelo probabilístico (por ejemplo, modelo de regulación codependiente y/o independiente) como se describe con mayor detalle a continuación. Debido al número potencialmente muy grande de posibles correlaciones de interacción, se contempla adicionalmente que el módulo de procesamiento ómico determinará un nivel de significancia de dependencia entre los parámetros reguladores (de un solo nodo) y la actividad de la ruta y/o el significado de la dependencia condicional entre los parámetros reguladores (de un solo nodo) dada una actividad de la ruta. De esa manera, se puede dar un enfoque analítico a las correlaciones de interacción con el significado estadísticamente más alto, como también se analiza con mayor detalle a continuación.

Aunque no se limitan al tema de la invención, los inventores también descubrieron que el análisis de las correlaciones de interacción y su significado se pueden refinar aún más mediante una manipulación estadística que determina el signo (positivo/activación, o negativo/inhibición) de la interacción para los parámetros reguladores. El uso de las correlaciones de interacción así determinadas y su influencia en la ruta ahora proporcionará una comprensión significativamente mejorada de las redes de rutas y el flujo de señales a través de dichas rutas.

Por lo tanto, y visto desde una perspectiva diferente, debe apreciarse que se puede generar un modelo de ruta obteniendo, a través de una interfaz de entrada ómica, al menos un conjunto de datos ómico (por ejemplo, datos del genoma completo, datos del genoma parcial o objetos de secuencia diferencial). Un módulo de procesamiento ómico accede entonces a un modelo de ruta (por ejemplo, previamente determinado) que tiene una pluralidad de elementos de ruta en los cuales al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores. El módulo de procesamiento ómico luego deduce, basado en el conjunto de datos ómico y/o el modelo de ruta, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores, y el modelo de ruta se actualiza posteriormente en función de las correlaciones de interacción.

Asimismo, debe reconocerse que mediante el uso de sistemas y métodos contemplados, se pueden identificar correlaciones de interacción específicas de subtipos para los parámetros reguladores de un nodo regulador en un modelo de ruta. Como antes, al menos un conjunto de datos ómico representativo de un tejido de subtipo se obtiene a través de una interfaz de entrada ómica, y un módulo de procesamiento ómico accede a un modelo de ruta determinado previamente. Las correlaciones de interacción de subtipo se derivan luego, a través del módulo de procesamiento ómico, del conjunto de datos ómico representativo del tejido de subtipo mediante análisis de probabilidad de interacciones entre la pluralidad de parámetros reguladores, como se explica con más detalle a continuación, y las correlaciones de interacción derivadas de subtipo se presentan (o incorporan) luego en el modelo de ruta. Si bien todas las clases de tipos secundarios de tejidos se consideran adecuados para su uso en el presente

documento, especialmente los subtipos contemplados incluyen tejido resistente al fármaco, tejido metastásico, tejido tratado con fármaco y/o una variante clonal de un tejido. Luego se pueden realizar experimentos de tipo experimental y/o teóricos (por ejemplo, *in vitro, in silico, in vivo*) para validar las correlaciones de interacción derivadas de subtipo. Por supuesto, y con respecto a los componentes y métodos de tales métodos, se aplican las mismas consideraciones que se proporcionaron anteriormente y a continuación.

Más específicamente, en el modelo gráfico probabilístico presentado en este documento, los estados de las moléculas biológicas (por ejemplo, proteínas, ARNm, complejos y biomoléculas pequeñas) de una muestra (por ejemplo, biopsia de tumor) se presentan como variables. Por ejemplo, para cada gen, las variables se utilizan para el número de copias del genoma de ese gen, el ARNm transcrito de ese gen, la proteína derivada de ese gen y, en la mayoría de los casos, una variable adicional no física que corresponde a la actividad biológica de un gen (como se anota en una ruta), que puede estar regulada por la modificación postraduccional de la proteína. También se pueden incluir variables que representan estados más abstractos, como la apoptosis, que comúnmente se anotan en las rutas.

Las interacciones causales que cambian el estado de las moléculas (por ejemplo, regulación de la transcripción génica, fosforilación de proteínas, formación de complejos) se representan como bordes dirigidos desde la variable reguladora hasta la variable regulada. Por lo tanto, para cada variable Y en el gráfico probabilístico del modelo, se introduce un factor en un modelo de probabilidad conjunta que relaciona el estado de la variable con el estado de todos sus reguladores: $F(Y|X_1,X_2,...,X_N)$, donde X_1 hasta X_N son las variables que regulan Y. Este factor es una tabla de probabilidad condicional: para cada configuración de Progenitores(Y), Σy F(Y = y|Progenitores(Y)) = 1. Observaciones de variables individuales, como el número de copia del genoma o la expresión del gen, se modelan como variables separadas, conectadas a la variable latente por un factor F(Y|X), también una tabla de probabilidad condicional. El estado de probabilidad conjunta total es entonces:

$$P(\Omega) = \frac{1}{Z}\prod_{Y\in\Omega} F(Y|\text{Progenitores}(Y))$$

donde Z es una constante de normalización requerida debido a los ciclos reguladores en la ruta.

Dadas las observaciones para una muestra, se puede resolver la distribución marginal de cada variable no observada, usando la implementación de propagación de creencias locas en libDAI con inferencia realizada en el espacio de probabilidad (en oposición al espacio logarítmico), una tolerancia de convergencia de 10^{-9} y con el programa de actualización de SEQFIX. Los parámetros para todas las funciones F se aprenden en un proceso de aprendizaje de máquina mediante la maximización de las expectativas en libDAI, deteniéndose cuando la proporción de probabilidades logarítmicas sucesivas es menor que 10^{-10}.

Debe apreciarse que los inventores ahora han introducido nuevas variables en el dogma central de cada gen que corresponden a los estados de transcripción, traducción y regulación de proteínas de cada gen, como se muestra en la Fig. 2A, que representa una estructura gráfica del factor típico. Este dogma central significa que cada gen codificador de proteína tendrá una estructura de dogma central idéntica y, por lo tanto, es posible compartir parámetros entre todos los genes. El programa regulador se modela luego en las variables de transcripción, traducción y regulación de proteínas para cada gen.

Modelos de regulación

El algoritmo desarrollado previamente (como se describe en los documentos WO 2013/062505 y WO 2011/139345) se amplió alterando la forma en que los algoritmos manejan los nodos de regulación. Para construir un gráfico de factores y permitir la comparación entre muchos tipos de datos, el algoritmo desarrollado previamente vuelve discretos los datos de entrada hacia abajo, hacia arriba o normal en relación con algún control. Los nodos de regulación recolectan señales de actividad de todos los genes involucrados en la regulación de un gen dado en algún punto a lo largo de la ruta del ADN hasta la proteína activa. Estas señales se recopilan en una sola variable que se conecta a la estructura del dogma central de un gen a través de un factor. Bajo el algoritmo desarrollado previamente, los nodos de regulación simplemente toman un voto de las señales entrantes para decidir si se transmitió una señal de activación o inhibición.

En contraste, en los sistemas y métodos de acuerdo con esta divulgación, la probabilidad de que cada ajuste de la variable Y hija que se pasa dado el ajuste de los nodos progenitores $X_1,...,X_N$ se aprende utilizando un proceso de aprendizaje de máquina. A continuación, se contrasta un modelo de regulación codependiente y uno independiente y se muestra como ejemplo en la Fig. 2B, que representa modelos de regulación alternativos para los nodos de transcripción, traducción y activación. En el modelo de regulación codependiente, se aprende una tabla de probabilidad condicional completa del hijo dados los progenitores, mientras que en el modelo de regulación independiente, se aprenden las probabilidades condicionales de los enlaces individuales y se utiliza un supuesto de Bayes sencillo para calcular la probabilidad del nodo hijo dados los progenitores.

Más específicamente, con el modelo de regulación codependiente de esta divulgación, la probabilidad se almacena directamente como un parámetro en una tabla de probabilidad condicional para todos los ajustes posibles de los progenitores y el hijo. En contraste, con el modelo de regulación independiente, P(Y) y $P(X_i|Y)$ se utilizan como

parámetros y el producto de los parámetros se calcula para encontrar la siguiente probabilidad:

$$F(Y|X_1, \ldots, X_N) = \frac{1}{Z} P(Y) \prod_i P(X_i|Y)$$

donde Z es una constante de normalización que corresponde a $P(X_1,...,X_N)$. Para inicializar los parámetros para el modelo de regulación independiente, P(Y) recibe una probabilidad igual hacia abajo, hacia arriba o normal, y la probabilidad inicial para $P(X_i|Y)$ se establece en función de la anotación del enlace en la ruta. Para enlaces marcados en la anotación como activadores P(abajo|abajo) = P(normal|normal) = P (arriba|arriba) = 0,8, y para inhibidores P (abajo| arriba) = P(normal|normal) = P(arriba|abajo) = 0,8 con todas las probabilidades de todos los demás ajustes fijados en 0,1. Las pruebas se realizaron utilizando una distribución uniforme en todos los ajustes para evaluar la importancia de utilizar este conocimiento previo de la ruta. El mismo procedimiento de escrutinio simple se usó como originalmente en el algoritmo desarrollado anteriormente como los parámetros iniciales para el aprendizaje de EM en el modelo de regulación codependiente de esta divulgación. Cuando □ = 0,001, se deduce que el 99,9% de la probabilidad se coloca en el estado de hijo que gana la aprobación y el 0,05% se coloca en los otros estados como las probabilidades iniciales.

Además, los inventores también permitieron la regulación de "activación" de complejos y familias de genes entre la proteína y los estados activos. Específicamente, cada familia y complejo ahora está modelado por un trío de variables: familia/complejo, regulación y activo, conectado con un solo factor F(activo|regulación, familia|complejo). Los reguladores de la familia o complejo están conectados a la variable activa, ya sea con el modelo de regulación codependiente de esta divulgación o el modelo de regulación independiente. Los componentes de la familia o el complejo están conectados a la variable de familia/complejo, utilizando un factor de ruido mínimo o de ruido máximo, con □ = 0,001. Por el contrario, solo se utilizó el factor de ruido mínimo o de ruido máximo en el algoritmo desarrollado previamente.

Estadísticas de regulación

Los inventores utilizaron pruebas de G para determinar el significado estadístico de la dependencia entre progenitores e hijos de los enlaces reguladores (primera ecuación), así como el significado estadístico de la dependencia condicional entre los progenitores dada una distribución de hijos (segunda ecuación):

$$G_{p-c} = 2 \sum_{i,j} O_{i,j} \ln \frac{O_{i,j}}{E_{i,j}}$$

$$= 2N \sum_{i,j} P(X_i, Y_j) \ln \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}$$

$$G_{p-p} = 2N \sum_{i,j} P(X_i, X_j|Y) \ln \frac{P(X_i, X_j|Y)}{P(X_i|Y)P(X_j|Y)}$$

Se debe tener en cuenta que la prueba G sigue la distribución X^2, de modo que se pueden encontrar valores P usando las distribuciones X^2 con 4 y 12 grados de libertad para la prueba progenitor-hijo y la prueba progenitor-progenitor, respectivamente. Los valores P se ajustan por la tasa de descubrimiento falsa (FDR) y los enlaces con P <0,05 ajustado se consideraron significativos. Aunque la prueba G (que es proporcional a la información mutua) es informativa sobre qué tan fuerte es una interacción, no proporciona detalles sobre el signo de la interacción (siendo la activación una interacción positiva y siendo la inhibición una interacción negativa).

Para obtener dicha información, los inventores calcularon tanto la correlación de Pearson entre el progenitor y el hijo, como la información mutua puntual ponderada, o WPMI (véase la fórmula a continuación) en todos los ajustes posibles del progenitor y el hijo. La correlación se calculó utilizando la distribución conjunta $P(X_i, Y) = P(X_i|Y)P(Y)$, y el significado se calculó utilizando la transformación de Fisher. La correlación entre dos progenitores dado el hijo también se calculó para determinar si los tres nodos formaban un ciclo de alimentación hacia adelante coherente o incoherente. Para comparar los resultados de la prueba G entre los grupos, se tomaron las diferencias de los rangos de la estadística G en cada grupo. El significado de esta estadística se calculó realizando una prueba de permutación con 5.000 permutaciones aleatorias de la membresía del grupo y luego ajustando para FDR. Para diferencias mayores que cualquiera de las observadas en las permutaciones, se usó el valor P más bajo posible como límite superior.

$$WPMI_{i,j} = P(X_i, Y_j) \ln \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}$$

Por lo tanto, debe reconocerse que el WPMI es simplemente cada elemento individual de la suma del puntaje G, y el vector de 9 valores de WPMI se puede organizar tan fácilmente como interpretar un mapa de calor. Los datos se pueden analizar utilizando un algoritmo de agrupación HOPACH (de Bioconductor), que intenta encontrar la cantidad

de agrupaciones que mejor se ajustan a los datos. Esto da como resultado diferentes números de agrupaciones para cada conjunto de IPL agrupadas. Para encontrar agrupaciones con un número consistente de agrupaciones entre todos los conjuntos de datos, los inventores colapsaron las agrupaciones más pequeñas al reasignar los miembros de agrupaciones pequeñas a la agrupación grande más cercana y las agrupaciones pequeñas colapsaron de esta manera para obtener una cantidad consistente de agrupaciones en todas las agrupaciones. Este método también sirvió para mantener los tamaños de las agrupaciones en todas las comparaciones.

Ejemplo

Existen numerosas maneras de producir un modelo de ruta, y se generó un modelo representativo a partir de Reactome, el PID y el análisis PID NCI de BioCarta, descargado en formato BioPAX Nivel 3 del 27 de febrero de 2012. Ese modelo de ruta comprendía 7.111 proteínas, 52 genes de ARN, 15 genes de miARN, 7.813 complejos, 1.574 familias de genes y 586 procesos biológicos abstractos. Hubo 8.603 interacciones que cambiaron el estado de activación de una molécula (inhibidor de 3.266), 2.120 enlaces de activación transcripcional, y 397 enlaces de represión transcripcional, y hubo 24.129 componentes para los 7.813 complejos, y 7.170 miembros de las 1.574 familias de genes.

Los inventores utilizaron DAVID para realizar el enriquecimiento de conjuntos de genes en los genes involucrados en las interacciones aprendidas por los sistemas y métodos de la invención. Para maximizar el número de genes reconocidos por DAVID, los complejos de genes y las familias se dividieron en sus genes componentes. El enriquecimiento para los genes involucrados en los enlaces se comparó con un antecedente de todos los genes en la ruta curada.

Una tabla de probabilidad condicional completa con N progenitores almacenará las probabilidades para todos los ajustes posibles 3^{N+1} de progenitores e hijos. Como algunos genes centrales en la ruta curada tienen más de 30 reguladores, el número de nodos progenitores que podrían unirse a un nodo hijo se limitó a 5 para evitar que el tamaño de estas tablas se vuelva prohibitivo. Para los genes regulados por más de cinco proteínas, se agregaron nodos intermedios a la gráfica para mantener este límite. Por lo tanto, un gen con 10 reguladores tendrá dos nodos intermedios con cinco reguladores unidos a cada nodo intermedio.

Utilizando un conjunto de datos de 1.936 muestras de tumor TCGA con datos de expresión génica y número de copias de 11 tipos de tejidos, se aprendieron las interacciones y las interacciones reguladoras, se determinó el significado de la interacción mediante una prueba G y se determinaron los signos de interacción con un valor de correlación como se describió anteriormente. De las 9.139 interacciones en el modelo de ruta que regula una proteína, se encontró que 7.631 (83,5%) eran significativas a un FDR de 0,05. Un análisis de componentes principales (PCA) de los vectores WPMI para cada interacción aprendida en toda la cohorte de TCGA reveló un gradiente de fuerte inhibición a fuerte activación. En la Fig. 3A-C se muestra un ejemplo de análisis del componente principal. Aquí, el panel (A) describe gráficamente el análisis del componente principal de los enlaces reguladores en la cohorte de TCGA en el que cada punto es la proyección de las 9 puntuaciones de WPMI para un enlace sobre los dos componentes principales. Las envolventes convexas muestran la membresía de la agrupación de medias k realizada en las puntuaciones de WPMI (no proyectadas), y los números de agrupación se colocan en el centroide de cada agrupación. El panel (B) ilustra la membresía del grupo de enlaces significativos etiquetados como activación e inhibición en la ruta, y el panel (C) muestra mapas de calor de los valores de WPMI de los centroides de los grupos que muestran un rango desde una fuerte inhibición (1) hasta una fuerte activación (5). La agrupación de medios k de los vectores WPMI encontró agrupaciones a lo largo de este gradiente que representan tipos de interacción canónica que van desde una fuerte activación hasta una fuerte inhibición. De 7.631 enlaces significativos, 78 (1%) se colocaron en un grupo donde el centroide iba en dirección opuesta a la forma en que se anotó el enlace en la ruta. La variedad de vectores de WPMI muestra que la EM fue capaz de aprender nuevos regímenes de interacción que parecen activadores e inhibidores, así como a patrones reguladores más complejos.

Usando medidas de correlación estadística (véase más arriba), los inventores luego evaluaron cada interacción como activación o inhibición y las compararon con el tipo de interacción anotado en el modelo de ruta. Hubo 7.357 enlaces tanto con correlación significativa como con puntajes de g y, de ellos, la correlación de 219 enlaces (3%) no estuvo de acuerdo con la dirección de la regulación en la ruta. Esto deja 7.138 (78%) enlaces que son significativos para ambas pruebas y concuerdan con los enlaces seleccionados. Los inventores también encontraron que algunos enlaces tenían altos valores de correlación pero poca significación a partir de nuestras pruebas g, que generalmente se observaba en los casos en que la distribución del progenitor o el hijo favorecía mucho a un solo estado.

De los enlaces aprendidos por el método de la invención, 1.197 tenían correlación significativa y puntajes g y no incluían complejos o familias. Para 51 de estos enlaces (4,3%), el signo del coeficiente de correlación no concuerda con la literatura. Por otro lado, al observar solo los perfiles de expresión génica, se encontraron 1.058 enlaces no familiares no complejos con una correlación significativa, pero 470 (44%) no estuvieron de acuerdo con el signo de la entrada de la ruta. Para una segunda comparación, los complejos y las familias se eliminaron en la ruta al conectar todos los genes que eran componentes de familias y complejos directamente a cualquier gen regulado por esas familias y complejos. Este procedimiento de aplanamiento resultó en 200.921 enlaces. Se encontró que 165.258 de estos enlaces tenían una correlación significativa entre los perfiles de expresión génica, y que 81.558 de los enlaces (49,4%) tenían una correlación que no estaba de acuerdo con la dirección del enlace en la ruta. Estos resultados

indican que los enlaces aprendidos por el método de la invención concuerdan significativamente mejor con la dirección de los enlaces en la literatura que con la correlación de los perfiles de expresión génica.

Ejecutando el PCA y el análisis de agrupación en solo las puntuaciones WPMI aprendidas de pacientes con cáncer de ovario TCGA (OV) (N = 416) y sin regulación de activación del complejo y familiar produjo resultados muy similares a los de PCA y centros de agrupación que se muestran en las Figs. 3A y 3C, pero encontró menos enlaces significativos y una mayor proporción de enlaces que se anotaron como activadores y se aprendieron como inhibidores o viceversa (Fig. 4A). Cuando se usó una inicialización plana de $P(X_i|Y) = 1/3$ (Fig. 4B), los inventores encontraron que los centros del agrupación se mapearon nuevamente en un gradiente desde la activación hasta la inhibición, y hubo menos enlaces significativos y una mayor proporción de desacuerdos en la dirección del enlace que con la configuración inicial que incluye información de la dirección.

Para probar el supuesto de independencia de Bayes sencillo presentado en la Fig. 2, los sistemas y métodos de acuerdo con el concepto inventivo se ejecutaron con modelos de regulación independientes y codependientes en las muestras de cáncer de ovario TCGA. Los inventores probaron el supuesto de independencia condicional en las expectativas calculadas en cada etapa EM de la ejecución (véase la Fig. 5A). La Fig. 5A ilustra el porcentaje de nodos hijos únicos que fallan en las siguientes pruebas en cada etapa de EM de un proceso que aprende una probabilidad condicional completa (Leyenda: i. una prueba del significado de la independencia condicional de cualquiera de los dos progenitores dado el hijo. ii. la prueba i y al menos uno de los progenitores que falla está significativamente vinculado al hijo. iii. la prueba i y el triplete que falla es incoherente, iv. las pruebas i, ii y iii. En cada etapa del aprendizaje, se encontraron menos correguladores que dependen el uno del otro. Debido a los pequeños bucles de retroalimentación en la ruta, tal como un factor de transcripción que regula su propia transcripción, se podría esperar que la suposición de independencia falle en algunos casos. Además, es bastante común para dos complejos muy similares, que difieren en una sola molécula, para corregular el mismo nodo hijo, en cuyo caso también se esperaría que la prueba de independencia condicional fallara, a pesar de que existe poco conflicto. Por consiguiente, los inventores dividen los casos donde dos correguladores fallan la prueba de independencia en clases 'coherentes' e 'incoherentes', como se muestra esquemáticamente en la Fig. 5B. La Fig. 5B ilustra esquemáticamente ejemplos de tripletes coherentes versus incoherentes. Las flechas corresponden a la correlación con una cabeza puntiaguda para una correlación positiva (activación) y una cabeza plana para una correlación negativa (inhibición). Las interacciones entre los progenitores no se encuentran en la literatura, por lo que se usaron flechas de doble sentido porque la dirección de esa interacción era desconocida.

Además, dos correguladores pueden fallar la prueba de independencia incluso si uno de los correguladores es un regulador insignificante, debido a la fuerza del otro regulador. Por lo tanto, los inventores también consideraron el subconjunto de casos donde ambos correguladores son significativos por sí mismos, y las pruebas muestran que los parámetros iniciales producidos por el método de aprobación ponderado hacen que casi el 50% de los nodos hijos no pasen la prueba de independencia condicional, pero como el algoritmo EM aprende más configuraciones de parámetros probables, cada vez menos nodos fallan en la prueba. La combinación de todas nuestras pruebas muestra que es probable que solo menos del 5% de los nodos hijos tengan reguladores codependientes de manera significativa.

Usando las muestras de cáncer de ovario, los inventores agruparon además las predicciones de actividad de la proteína producidas por el algoritmo desarrollado previamente (véanse los documentos WO 2013/062505 y WO 2011/139345) y aquellas de los modelos de regulación codependientes e independientes. Luego se realizó un análisis de Kaplan-Meier en estas agrupaciones para ver si tenían perfiles de supervivencia significativamente diferentes (Fig. 6). Aquí, se muestran las curvas de supervivencia de Kaplan-Meier de 416 pacientes en la cohorte ovárica TCGA agrupada mediante la actividad de la ruta integrada utilizando (Fig. 6A) el algoritmo desarrollado previamente, (Fig. 6B) el algoritmo inventivo que aprende las tablas de probabilidad condicional completas de los nodos reguladores, y (Fig. 6C) el algoritmo de la invención que aprende la probabilidad condicional de enlaces simples y el uso de un supuesto de Bayes sencillo. Los inventores encontraron que las agrupaciones producidas utilizando predicciones de actividad de modelos de regulación independientes fueron los más separables por su supervivencia (rango logarítmico $P= 2{,}0 \times 10^{-4}$). Los inventores también realizaron esta prueba utilizando el modelo de regulación independiente con una configuración inicial plana para los parámetros $P(X_i|Y)$ y encontraron que se desempeñó peor que el algoritmo desarrollado previamente. Nuevamente, esto indica que el método de aprendizaje requiere un conocimiento previo sobre el tipo de interacción que se pierde cuando se utiliza una configuración de interacción inicial plana.

La Fig. 7 muestra el uso del enlace diferencial de tejido en el más significativo al colorear cada interacción por su puntuación de correlación en un tejido y establecer su saturación de manera proporcional a su importancia. Se observaron las puntuaciones g diferenciales más fuertes para los enlaces regulados por los genes y complejos clave del cáncer, incluidos TP53, MYC/MAX, HIF1A/ARNT, TAp73a, E2F1 y PPARA-RXRA. De particular interés son los enlaces regulados por PPARA-RXRA principalmente diferentes dentro de GBM [cerebro y KIRC (riñón)] y los enlaces reguladores de TAp73a en OV (ovario) y en menor grado en UCEC (endometrio uterino). Las Figuras 8A y 8B muestran una gráfica de las señales de WPMI agrupadas por tejido para los enlaces de activación de PPARA-RXRA y TAp73a, donde se encuentran pesos significativamente mayores en la diagonal de activación, lo que indica un mayor uso de estos enlaces como activadores en esos tejidos. Como se puede ver en la Fig. 8A que muestra los valores de WPMI para los enlaces con PPARA:RXRA como nodo progenitor, hay una señal de activación más fuerte en GBM y KIRC, mientras que la Fig. 8B muestra los valores de WPMI para los enlaces con TAp73a como nodo progenitor, lo que indica activación en VO.

La firma de la actividad de TAp73 indica potencialmente un patrón reproductivo u hormonal femenino de patogénesis asociado con la expresión de p73. TAp73 promueve la expresión de inhibidores del ciclo celular e inductores de apoptosis, uno de los cuales es el supresor de tumores BAX, que actúa como un inhibidor de la actividad del oncogén BCL2. Se sabe que BCL2 es altamente expresado en el cáncer de ovario seroso, y los resultados aquí muestran que aunque TAp73 es altamente expresado y es un fuerte promotor de la expresión de BAX (y por lo tanto la inhibición de BCL2), no obstante, es inefectivo para retardar la tumorigénesis, lo que sugiere que la inhibición de la molécula pequeña de BCL2 puede ser igualmente ineficaz. No es sorprendente que los tratamientos con un sólo agente del cáncer de ovario con inhibidores de molécula pequeña de BCL2, a pesar de la alta expresión de BCL2 en el cáncer de ovario seroso, no hayan tenido éxito hasta la fecha, lo que sugiere un bloqueo o atenuación de la actividad mediada por TAp73 en este tipo de cáncer. Es importante tener en cuenta que casi todas las muestras de ovario serosas aquí presentan mutaciones en p53, lo que quizás sugiera una derivación hacia arriba de la tumorigénesis que quizás supere la sobreexpresión de TAp73 o el aumento de la actividad. Otros grupos también han demostrado la importancia de la actividad de PPARA-RXRA en GBM y KIRC y su sensibilidad al fenofibrato, un agonista de PPARA. Las señales específicas de tejido identificadas a través de este análisis parecen reiterar descubrimientos biológicos recientes que parecen ser únicos cuando se examinan en el contexto del conjunto de datos actual de TCGA.

Los enlaces más significativos aprendidos a través de toda la cohorte de TCGA (véase la Tabla 1) son varios genes de cáncer conocidos que incluyen el factor de transcripción A1 de la caja cabeza de horquilla, p53 y el receptor alfa de estrógeno. Para realizar un enriquecimiento de conjuntos de genes con DAVID en los genes involucrados en las 50 interacciones con las puntuaciones G más altas, los inventores reemplazaron las familias y los complejos con sus genes componentes. Esto produjo 112 genes únicos que fueron reconocidos por DAVID a partir de los 50 enlaces principales. Se encontró que estos genes estaban significativamente enriquecidos ($P < 1e^{-7}$) para una serie de términos KEGG relevantes que incluyen "rutas en el cáncer", "apoptosis", "ruta de señalización Jak-STAT" y "ruta de señalización MAPK" como una serie de diferentes términos específicos del tipo de cáncer. Luego, los inventores compararon este resultado con lo que podría encontrarse al observar solo la correlación de la expresión génica de los genes que están enlazados en la ruta. Los inventores necesitaron tomar los 200 pares de expresión génica principales por la correlación de Pearson de la ruta aplanada para obtener un conjunto de genes únicos de tamaño comparable (N = 119) al conjunto producido por el algoritmo de la invención. Aunque ambos conjuntos de genes produjeron enriquecimientos similares para los términos de Ontología de Genes para procesos biológicos (GOTERM_BP_FAT), se encontraron muchos menos términos KEGG mediante el uso de la correlación de la expresión génica que mediante los enlaces aprendidos (20 versus 46 en FDR <0,05) y el FDR. Los términos KEGG que se superponían entre los dos conjuntos tenían un FDR más bajo en el conjunto determinado. Para asegurarse de que el aplanamiento de familias y complejos en la ruta no influyera en estos resultados, los inventores repitieron este análisis para enlaces no familiares y no complejos solo en la ruta y encontraron resultados similares (se encontraron 20 términos KEGG para enlaces aprendidos versus 3 para la correlación de la expresión en FDR <0,05).

Tabla 1. Vínculos reguladores con la puntuación de prueba g más alta en toda la cohorte de TCGA

Progenitor	Hijo	Puntuación g	Dirección
FOXA1	SFTPA (familia):txreg	3247,197	↑
HNF1A	HNF4A (familia):txreg	3208,440	↑
GATA1	Globina alfa (familia):txreg	3065,885	↑
ONECUT1	HNF1B (familia):txreg	3008,945	↑
Tetrámero p53 (complejo)	MDM2:txreg[a]	2931,148	↑
KLF4	Preprogherina (familia):txreg	2914,620	↑
PDX1	NR5A2 (familia):txreg	2872,275	↑
Tetrámero p53 (complejo)	SFN:txreg[a]	2811,958	↑
Homodímero ER alfa (complejo)	Tubulina alfa (familia):txreg	2781,369	↑
FOXM1	CENPA:txreg	2739,028	↑

Los valores de p para todos los enlaces son menores que 1e-323.

[a]Nodo intermedio.

Los inventores también compararon la fuerza de los enlaces entre los subtipos de cáncer de mama para obtener una idea de las diferencias reguladoras entre los subtipos (véase la Tabla 2). Esta comparación, así como otras comparaciones entre tejidos, nunca encontraron enlaces que cambiaran completamente la dirección de activación a

inhibición. En su lugar, los inventores a menudo observaron que los enlaces se apagaron o encendieron (por ejemplo, cambiaron de un activador fuerte a neutral). Debido a que la dirección rara vez cambia, a los inventores les pareció informativo simplemente observar las diferencias entre el significado de la puntuación G de los enlaces. Los inventores utilizaron la diferencia de rango de las puntuaciones G para comparar entre grupos a fin de ajustar la dependencia de la puntuación G en el tamaño de la muestra. Muchos de los enlaces con las diferencias de rango más altas tenían los mismos progenitores. Por ese motivo, la Tabla 2 muestra los enlaces con la diferencia de rango más alta con base en el progenitor. En 9 de los 10 enlaces principales que eran más fuertes en los tumores basales, HIF1A era el progenitor y los cuatro enlaces principales más fuertes en los tumores Luminal A tenían CEBPB como un progenitor.

Tabla 2. Enlaces reguladores con P ajustado <0,05 en los tumores de cáncer de mama Basal (N = 92) o Luminal A (N = 218), y las diferencias de rango más altas en las puntuaciones G por progenitor.

Progenitor	Hijo	Valor P basal	Valor P luminal	Diferencia de rango	Dirección
HIF1A/ARNT	HK1	1,61e-3	0,834	7826	↑
(complejo)	RRM1	9,20e-3	0,854	7632	↑
E2F3/DP,TFE3	PPP3CA	3,09e-2	0,493	5203	↑
(complejo)	WAF1	3,48e-2	0,459	4924	↑
MYB	TP73	6,59e-3	0,343	4225	↑
E2F1(DP)	HSP90B1	0,879	9,65e-3	6275	↑
(complejo)	AChR (familia)	0,833	0,0256	4742	↑
E2F1/DP/PCAF	CDKN2C	0,771	5,94e-4	4700	Insignificante
(complejo)	SERPINB5	0,808	0,0300	4264	↑
CEBPB	LEF1	0,775	9,18e-3	4250	↑
JUN					
SP1					
Daño del ADN					
(resumen)					
LEF1/Catenina beta/PITX2					
(complejo)					
Nota: El P ajustado de todas las diferencias de rango en esta tabla fue < 4,8e-4. Todos los bordes fueron anotados como activadores transcripcionales. La tabla completa es material suplementario.					

Para identificar las actividades clínicamente relevantes y las fortalezas de los enlaces, los inventores examinaron a los pacientes con cáncer de mama con receptores de estrógeno positivos (ER+) y realizaron una regresión Cox regularizada de CGA de los datos de supervivencia de la TCGA en ambas puntuaciones g del enlace e IPL para identificar el número óptimo de características para dividir mejor la cohorte. En la lambda mínima, el modelo coxnet contenía nueve características que dividían mejor a los pacientes cáncer de mama con ER+ (véase la Tabla 3). Cuatro de las nueve características fueron puntuaciones g de enlace, que ilustran la utilidad independiente de estas puntuaciones como posibles marcadores de pronóstico.

Tabla 3. Características de la ruta (bordes y nodos) asociadas con la supervivencia en pacientes de cáncer de mama con ER+

Característica	Coeficiente de riesgo de Cox
GLI2A → GLI1	0,08484
HIF1A/ARNT (complejo) → CP	0,07835

MYB → CEBPB	0,00462
E2F1/DP (complejo) → SIRTI	-0,00072
p300/CBP (complejo)	-0,00204
SDC3	-0,04840
p300/CBP/RELA/p50 (complejo)	-0,11126
TAp73a (tetrámero) (complejo)	-0,11301
TCF1E/Catenina beta (complejo)	-0,16129

Nota: Los bordes se identifican con →, y todos los bordes encontrados se anotan como activadores transcripcionales en la ruta.

CEBPB y HIF1A/ARNT aparecieron en ambas Tablas 2 y 3. CEBPB es un factor de transcripción que se ha asociado con la progresión tumoral, mal pronóstico y estado ER negativo. Además, la sobreexpresión de HSP90B1, una proteína de choque térmico regulada por CEBPB y que se encuentra en la Tabla 2, se ha asociado con metástasis distante y disminución de la supervivencia general en pacientes con cáncer de mama con buenos pronósticos. HSP90B1 se ha sometido a ensayos clínicos como inmunoterapia para el melanoma con el nombre de vitespen. La sobreexpresión de HIF1A/ARNT es clínicamente relevante en el cáncer de mama ER- y PR-, donde las variantes de empalme se han asociado con una supervivencia reducida sin metástasis. Debido a que los tumores basales son generalmente ER-, y los tumores Luminal A son generalmente ER+, la fuerza del enlace diferencial podría deberse a un aumento en la aparición de la variante de empalme en los tumores basales. Los dos enlaces principales por diferencia de rango del puntaje G entre basal y luminal son HIF1A/ARNT que activan HK1 y HK2 (hexoquinasas), HK2 participa en el metabolismo de la glucosa y la apoptosis, y se ha asociado con metástasis cerebrales de cáncer de mama y con escasa supervivencia posterior a craneotomía. Estos hallazgos indican la posibilidad de encontrar enlaces que sean relevantes al contrastar entre subtipos de tumores y al buscar enlaces dentro de un subtipo que sean predictivos de una variable clínica.

Con base en lo anterior, debe apreciarse que los sistemas y métodos contemplados permiten una combinación de datos ómicos múltiples para conocer la fuerza y el signo de las interacciones reguladoras seleccionadas a partir de la literatura. El supuesto de independencia condicional permite una reducción en la complejidad del modelo y permite una estimación eficiente de los parámetros reguladores utilizando los conjuntos de datos existentes. Además, los inventores también demostraron que el supuesto de independencia es válido para la gran mayoría de los programas de regulación celular. Además, cuando el supuesto de independencia no se cumple, se contempla que los factores independientes podrían reemplazarse por factores más complejos que modelan adecuadamente un programa de regulación codependiente. Cuando se aplican estos parámetros aprendidos, se puede obtener una visión biológica simplemente observando los enlaces más fuertes en una cohorte de muestras u observando cómo cambian las interacciones entre los fenotipos de interés.

También debe apreciarse que aunque los subtipos de cáncer usan diferentes interacciones, una interacción generalmente tiene un signo consistente cuando se usa en un tumor particular. Aún más, la concordancia del signo de interacción aprendido y el signo de interacción en las bases de datos, a pesar de las diversas formas en que el signo de interacción se anota en el lenguaje BioPAX a través de las bases de datos de rutas, indica que las bases de datos de rutas ya han catalogado con éxito y fidelidad miles de experimentos de WetLab en la literatura.

Además, debe apreciarse que la independencia de los correguladores proporciona beneficios computacionales para la inferencia del modelo y el aprendizaje de parámetros, y también ayuda en la interpretación del modelo. La capacidad de ser factorizados de los modelos de regulación corresponde a la linealidad logarítmica. Sin embargo, un gran número de reguladores en el modelo son complejos, y el factor de formación del complejo es una función de ruido máximo no lineal. Por lo tanto, la no linealidad de la regulación todavía puede codificarse en el gráfico de factores que representan complejos físicos. Esto otorga plausibilidad a una interpretación física de la mayoría de los enlaces de regulación en la ruta: la unión competitiva de reguladores independientes debe combinarse linealmente, siempre que las entidades físicas verdaderamente independientes hayan sido capturadas como complejos. Si esta interpretación física es cierta, entonces debería haber una correspondencia entre las fortalezas relativas de las constantes de unión físicas medidas y las puntuaciones de interacción determinadas. En los casos en que el supuesto de independencia no se cumple, es probable que exista un cofactor latente, que podría modelarse reemplazando $P(Y|X_1)P(Y|X_2)$ con un factor como $P(Y|X_1,X_2)$.

Como los métodos y sistemas contemplados son capaces de diferenciar las correlaciones de interacción entre subtipos de tejido, los inventores también contemplan un método para clasificar un conjunto de datos ómico representativo de un tejido (por ejemplo, obtenido de una biopsia de tumor) como perteneciente a un subtipo de tejido específico (por ejemplo, como perteneciente a un tumor resistente al tratamiento con respecto a un fármaco en particular). De manera

similar a los métodos discutidos anteriormente, los métodos contemplados obtendrán primero a través de una interfaz de entrada ómica el conjunto de datos ómico representativo del tejido, y luego derivarán, para el conjunto de datos ómico, un conjunto de correlaciones de interacción entre una pluralidad de parámetros reguladores de un nodo regulador en un modelo de ruta. El conjunto así derivado de correlaciones de interacción se empareja luego con un conjunto de correlaciones de interacción previamente conocido que está asociado con un tejido específico del subtipo conocido, y cuando se desea, el emparejamiento se usa luego para la clasificación del conjunto de datos ómico (por ejemplo, para ser representativo del tejido específico del subtipo conocido, y con eso para clasificar el tejido como perteneciente al subtipo). Por lo tanto, debe apreciarse que los sistemas y métodos contemplados permitirán la caracterización de un tejido en términos de un subtipo simplemente basado en una o más firmas de correlación de interacción. Entre otros subtipos de tejido contemplados, los subtipos especialmente ventajosos incluyen tejido resistente a fármaco, tejido metastásico, tejido tratado con fármaco o una variante clonal de un tejido.

Además, como los sistemas y métodos contemplados permiten la identificación del flujo de señal a través de una ruta de señalización y/o una red de rutas, debe apreciarse que los sistemas y métodos contemplados también serán útiles para identificar un objetivo que se pueda tratar con fármaco en un modelo de ruta. Dicha identificación normalmente incluirá etapas de (a) obtener, a través de una interfaz de entrada ómica, un conjunto de datos ómico representativo de un tejido, (b) derivar, para el conjunto de datos ómico, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores del nodo regulador en el modelo de ruta, y (c) la identificación de un fármaco que afecta la actividad de la ruta en la que se predice que el fármaco interfiere con las correlaciones de interacción. Más típicamente, el nodo regulador afecta al menos a uno de transcripción, traducción y modificación postraduccional de una proteína, y el fármaco es un fármaco disponible comercialmente y tiene un modo de acción conocido.

De este modo, como se conocen las correlaciones de interacción específicas entre los parámetros reguladores de una ruta, la ruta objetivo en un modelo de ruta ahora se puede identificar fácilmente usando un conjunto de datos ómico representativo de un tejido, y una derivación, para el conjunto de datos ómico, de un conjunto de correlaciones de interacción entre los parámetros reguladores de un nodo regulador en un modelo de ruta. Cuando un fármaco tiene un efecto conocido en la correlación de interacción, el fármaco se puede utilizar para dirigirse a la ruta objetivo. Por ejemplo, el efecto conocido de un fármaco puede ser un efecto inhibitorio sobre una quinasa, un efecto inhibidor sobre un receptor y un efecto inhibidor sobre la transcripción. Por lo tanto, y entre otras rutas objetivo adecuadas, las rutas objetivos especialmente contempladas incluyen una ruta regulada por calcio/calmodulina, una ruta de citoquina, una ruta de quimioquina, una ruta regulada por el factor de crecimiento, una ruta regulada por hormona, una ruta regulada por MAP quinasa, una ruta regulada por fosfatasa, y una ruta regulada por Ras. Dependiendo del resultado del análisis de la ruta, los consejos de tratamiento pueden basarse en la ruta identificada.

Además, debe apreciarse que el tratamiento no necesita realizarse realmente en un paciente, sino puede simularse una vez que se conocen una o más correlaciones de interacción específicas entre los parámetros reguladores de una ruta. Dicha simulación se puede usar para predecir el resultado del tratamiento o la identificación de múltiples fármacos para detectar señales efectivamente bajas a través de las rutas. Por lo tanto, los métodos contemplados también incluirán un método para simulación *in silico* del efecto de tratamiento de un fármaco que incluye una etapa para obtener un modelo de ruta que tiene una pluralidad de elementos de ruta en los cuales al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta como una función de una pluralidad de parámetros reguladores. Los métodos contemplados incluirán además una etapa para identificar un fármaco que se sabe que afecta al menos a un parámetro regulador, y otra etapa para alterar *in silico*, a través de un módulo de procesamiento ómico y basado en el efecto conocido del fármaco, al menos uno del nodo regulador, la actividad y al menos de los parámetros reguladores en el modelo de ruta, y aún otra etapa para determinar un efecto secundario de la alteración en el modelo de ruta. En general, el efecto secundario se encuentra en otro nodo regulador, otra actividad y otro parámetro regulador en el modelo de ruta.

Para los expertos en la técnica, debería ser evidente que son posibles muchas más modificaciones, además de las ya descritas, sin apartarse de los conceptos de la presente invención. Además, al interpretar tanto la especificación como las reivindicaciones, todos los términos deben interpretarse de la manera más amplia posible y coherente con el contexto. En particular, los términos "comprende" y "que comprende" deben interpretarse en referencia 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 están expresamente referenciados. Cuando las reivindicaciones de la especificación se refieren a al menos uno de los elementos seleccionados del grupo que consiste en A, B, C... y N, el texto debe interpretarse como que solo requiere un elemento del grupo, no A más N o B más N, etc.

REIVINDICACIONES

1. Un método implementado por ordenador para clasificar un tejido como perteneciente a un tejido específico de subtipo, que comprende:

(a) obtener, a través de un módulo de interfaz de entrada ómico (120), al menos un conjunto de datos ómico (135) representativo del tejido;

(b) acceder, a través de un módulo de procesamiento ómico (170), a un modelo de ruta biológica (150) que tiene una pluralidad de elementos de la ruta que comprenden al menos uno de una secuencia de ADN, una secuencia de ARN, una proteína y una función de proteína, en la que al menos dos de los elementos están acoplados entre sí a través de una ruta que tiene un nodo regulador que controla la actividad a lo largo de la ruta en función de una pluralidad de parámetros reguladores;

i) cuando el elemento de la ruta comprende una secuencia de ADN, al menos uno de la pluralidad de parámetros reguladores se selecciona del grupo que consiste en un factor de transcripción, un activador de la transcripción, una subunidad de ARN polimerasa, un elemento regulador en cis, un elemento regulador en trans, una histona acetilada, una histona metilada y un represor,

ii) cuando el elemento de la ruta comprende una secuencia de ARN, al menos uno de la pluralidad de parámetros reguladores se selecciona del grupo que consiste en un factor de iniciación, un factor de traducción, una proteína de unión a ARN, una proteína ribosómica, un ARNpi y una proteína de unión a poliA, y

iii) cuando el elemento de la ruta comprende una proteína, al menos uno de la pluralidad de parámetros reguladores es una fosforilación, una acilación, una escisión proteolítica y asociación con al menos una segunda proteína;

(c) inferir, mediante el módulo (170) de procesamiento ómico, basado en al menos un conjunto de datos (135) ómico y el modelo (150) de ruta, un conjunto de correlaciones de interacción entre la pluralidad de parámetros reguladores, en el que las probabilidades condicionales de enlaces individuales se aprenden y se utiliza una suposición de Bayes sencilla para calcular la probabilidad de un nodo hijo Y dados los progenitores X_1, ... X_n, en donde la probabilidad F se calcula con base en la expresión:

$$F(Y|X_1, \ldots, X_N) = \frac{1}{Z} P(Y) \prod_i P(X_i|Y)$$

en la que Z es una constante de normalización que corresponde a $P(X_1,...X_n)$, siendo el modelo de ruta un modelo probabilístico configurado para usar gráficos de factores usando un modelo de regulación independiente;

(d) actualizar el modelo de ruta (150) basado en las correlaciones de interacción aprendidas, en el que la interfaz de entrada ómica es una interfaz de computación configurada para recibir uno o más conjuntos de datos ómicos, y en el que el módulo de procesamiento ómico es una parte de un dispositivo de computación, en el que una prueba G determina la significación estadística de la dependencia entre los progenitores que proporcionan una distribución de hijos, y

en el que una correlación de Pearson o información mutua puntual ponderada (WPMI) determina el signo de interacción para los parámetros reguladores,

(e) hacer coincidir el conjunto derivado de correlaciones de interacción con un conjunto conocido a priori de correlaciones de interacción que está asociado con un tejido específico de subtipo conocido; y

(f) utilizar el emparejamiento para clasificar que el conjunto de datos ómico representativo del tejido pertenece al tejido específico de subtipo conocido, en el que el subtipo incluye tejido resistente al fármaco, tejido metastático, tejido tratado con fármaco o una variante clonal de un tejido.

2. El método implementado por ordenador de la reivindicación 1, en el que los conjuntos de datos (135) ómicos comprenden datos de todo el genoma, datos parciales del genoma u objetos de secuencia diferencial, y en el que los conjuntos de datos (135) ómicos se obtienen de una base de datos (130) genómica, un servidor (130) BAM, o un dispositivo (130) de secuenciación.

3. Aparato configurado para llevar a cabo el método de la reivindicación 1 o la reivindicación 2.

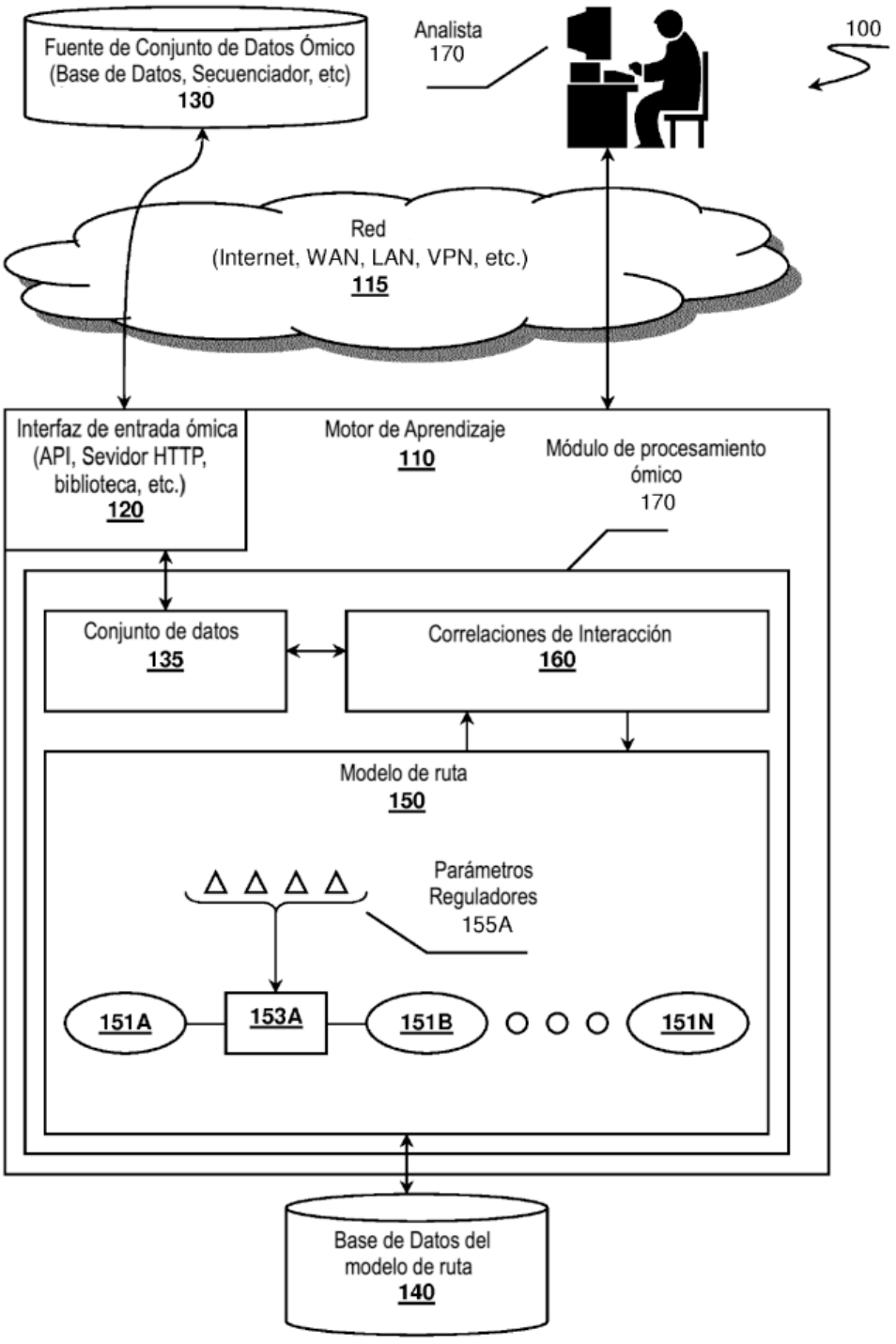

Fuente de Conjunto de Datos Ómico
(Base de Datos, Secuenciador, etc)
130
Analista
170
100
Red
(Internet, WAN, LAN, VPN, etc.)
115
Interfaz de entrada ómica
(API, Sevidor HTTP,
biblioteca, etc.)
120
Motor de Aprendizaje
110
Módulo de procesamiento
ómico
170
Conjunto de datos
135
Correlaciones de Interacción
160
Modelo de ruta
150
Parámetros
Reguladores
155A
151A
153A
151B
151N
Base de Datos del
modelo de ruta
140

Fig. 1

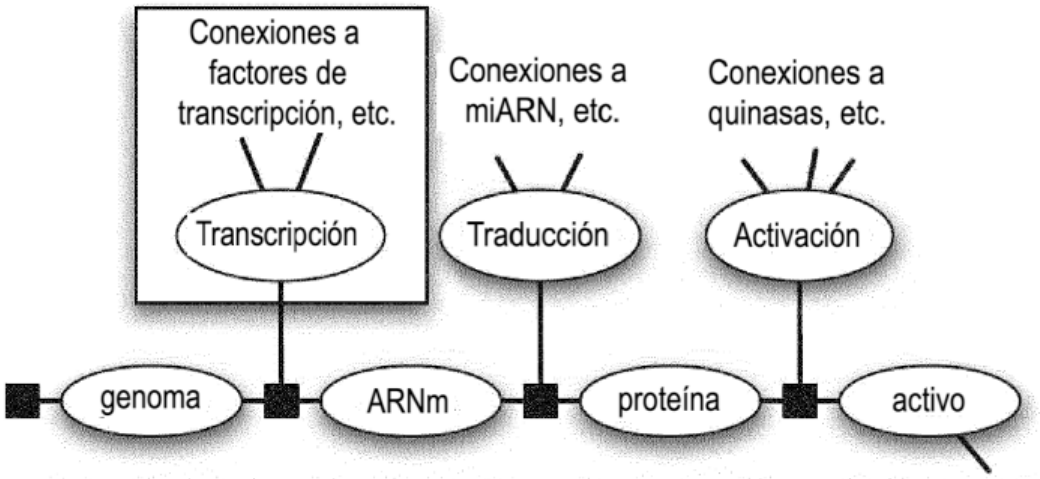

Conexiones a factores de transcripción, etc.
Conexiones a miARN, etc.
Conexiones a quinasas, etc.
Transcripción
Traducción
Activación
genoma
ARNm
proteína
activo

Fig. 2A

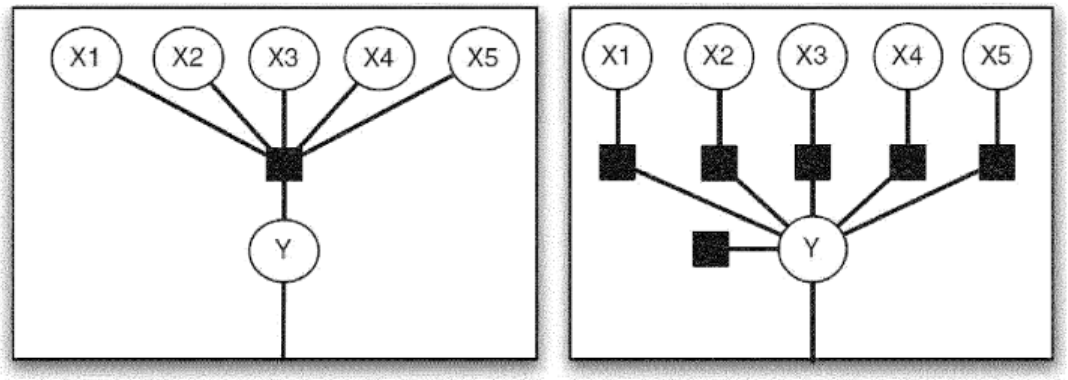

X1
X2
X3
X4
X5
Y
X1
X2
X3
X4
X5
Y
Modelo de Regulación Codependiente
Modelo de Regulación Independiente

Fig. 2B

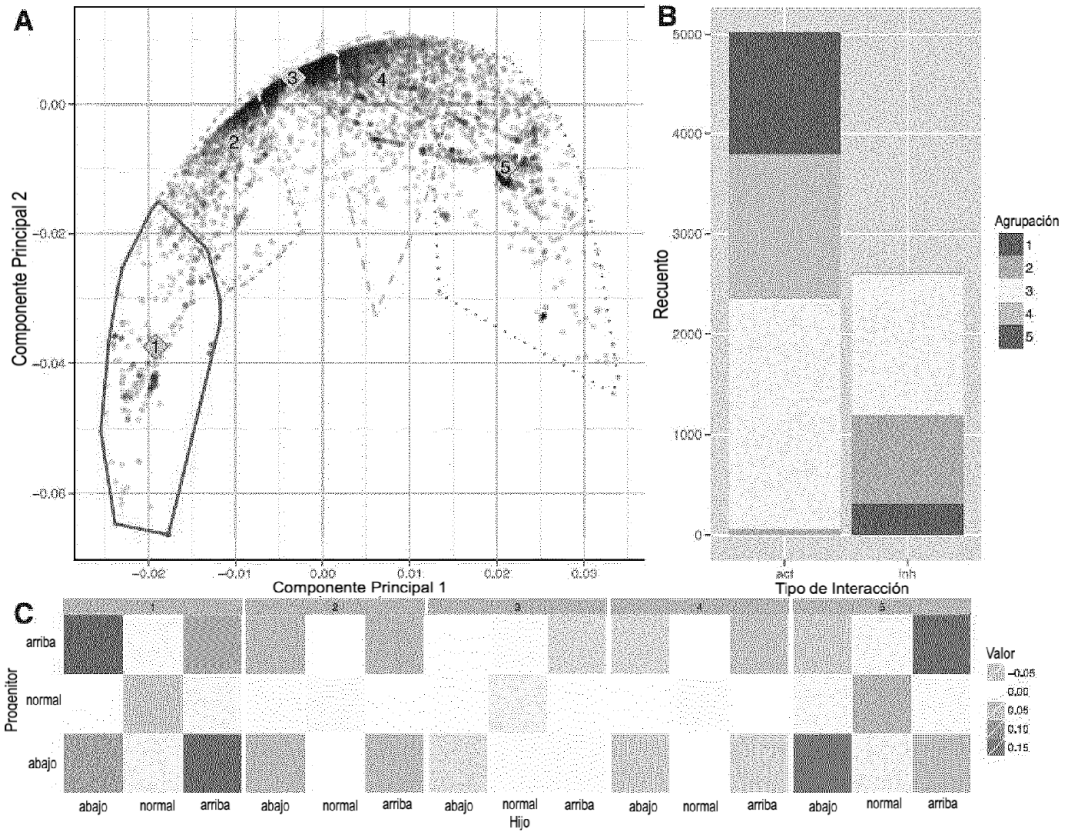

A
B
C
Componente Principal 2
Componente Principal 1
Recuento
Tipo de Interacción
act
inh
Agrupación
1
2
3
4
5
Progenitor
arriba
normal
abajo
Hijo
Valor

Fig. 3

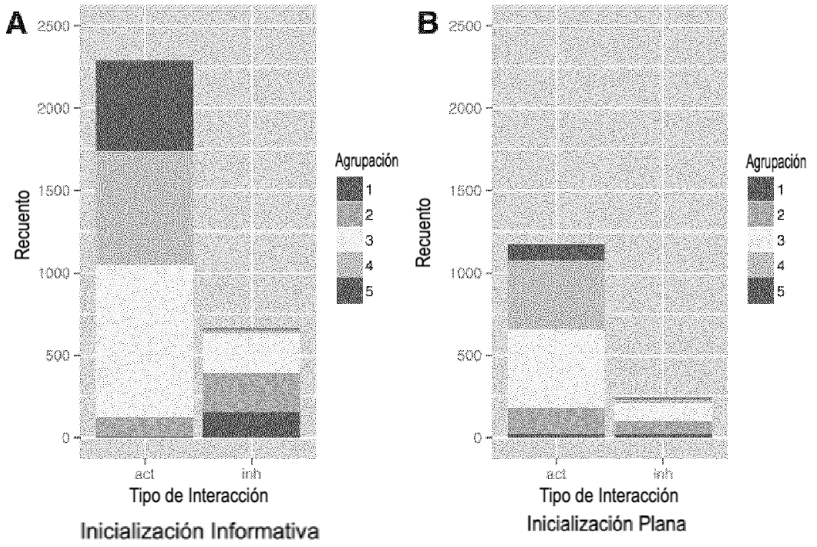

A
B
Recuento
2500
2000
1500
1000
500
0
act
inh
Tipo de Interacción
Agrupación
1
2
3
4
5
Inicialización Informativa
Inicialización Plana

Fig. 4

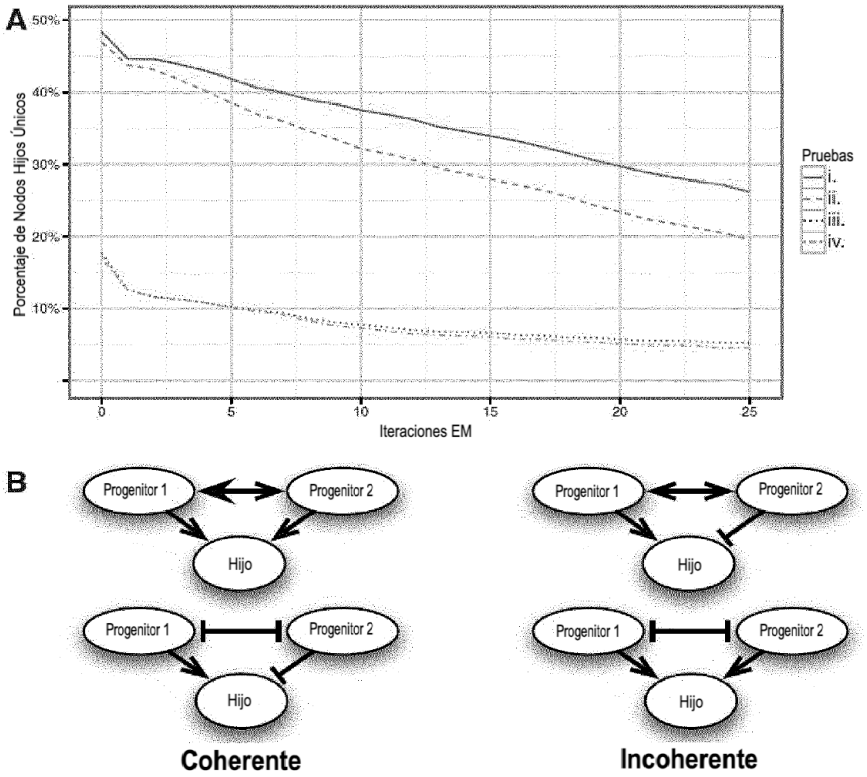
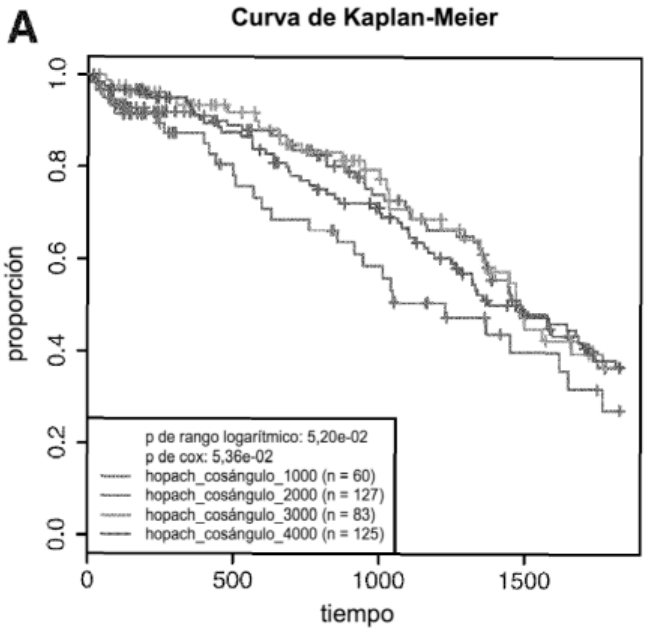
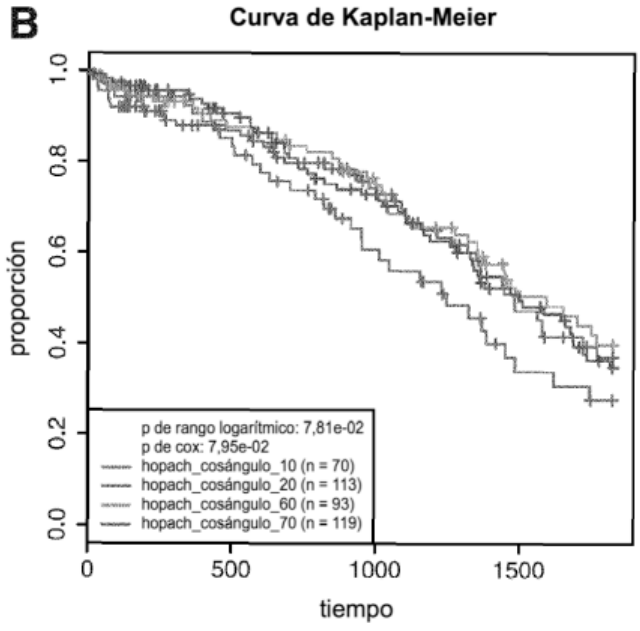

A
B
Porcentaje de Nodos Hijos Únicos
50%
40%
30%
20%
10%
0
5
10
15
20
25
Iteraciones EM
Pruebas
i.
ii.
iii.
iv.
Progenitor 1
Progenitor 2
Hijo
Coherente
Incoherente

Fig. 5



Modelo de Regulación Original

Fig. 6A



Modelo de Regulación Codependiente

Fig. 6B

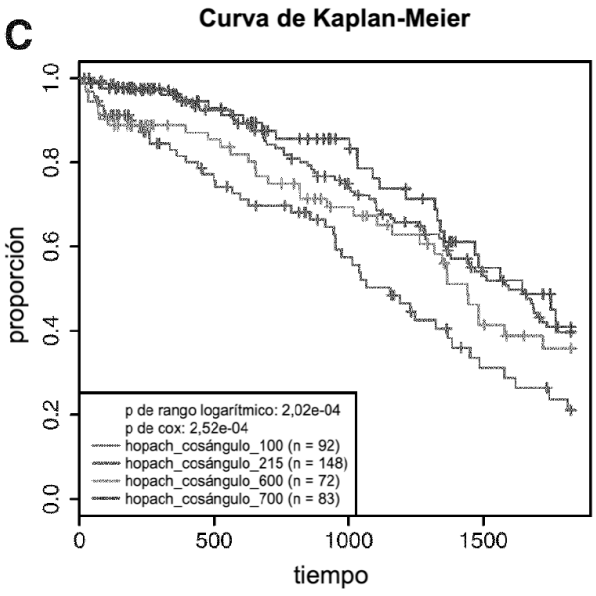

C
Curva de Kaplan-Meier
proporción
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
tiempo
p de rango logarítmico: 2,02e-04
p de cox: 2,52e-04
hopach_cosángulo_100 (n = 92)
hopach_cosángulo_215 (n = 148)
hopach_cosángulo_600 (n = 72)
hopach_cosángulo_700 (n = 83)
Modelo de Regulación Independiente

Fig. 6C

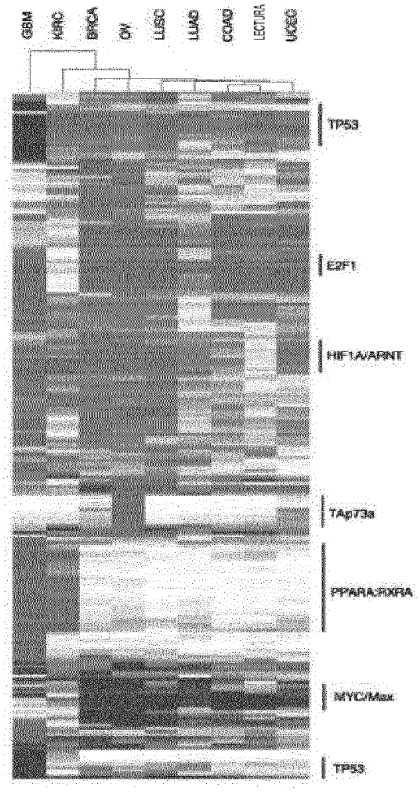

GBM
KIRC
BRCA
OV
LUSC
LUAD
COAD
LGG
UCEC
TP53
E2F1
HIF1A/ARNT
TAp73a
PPARA/RXRA
MYC/Max
TP53

Fig. 7

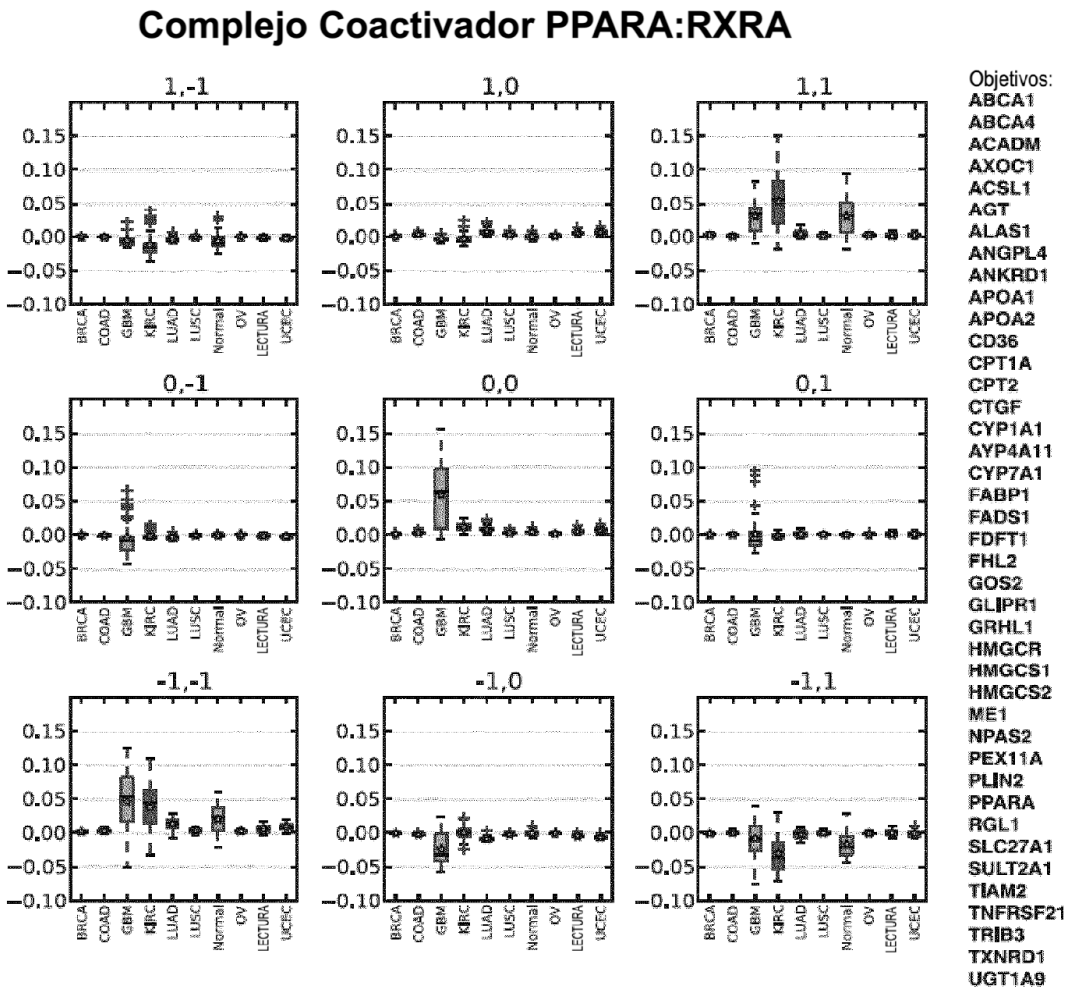

Complejo Coactivador PPARA:RXRA
1,-1
1,0
1,1
0,-1
0,0
0,1
-1,-1
-1,0
-1,1
0.15
0.10
0.05
0.00
−0.05
−0.10
BRCA
COAD
GBM
KIRC
LUAD
LUSC
Normal
OV
LECTURA
UCEC
Objetivos:
ABCA1
ABCA4
ACADM
AXOC1
ACSL1
AGT
ALAS1
ANGPL4
ANKRD1
APOA1
APOA2
CD36
CPT1A
CPT2
CTGF
CYP1A1
AYP4A11
CYP7A1
FABP1
FADS1
FDFT1
FHL2
GOS2
GLIPR1
GRHL1
HMGCR
HMGCS1
HMGCS2
ME1
NPAS2
PEX11A
PLIN2
PPARA
RGL1
SLC27A1
SULT2A1
TIAM2
TNFRSF21
TRIB3
TXNRD1
UGT1A9

Fig. 8A

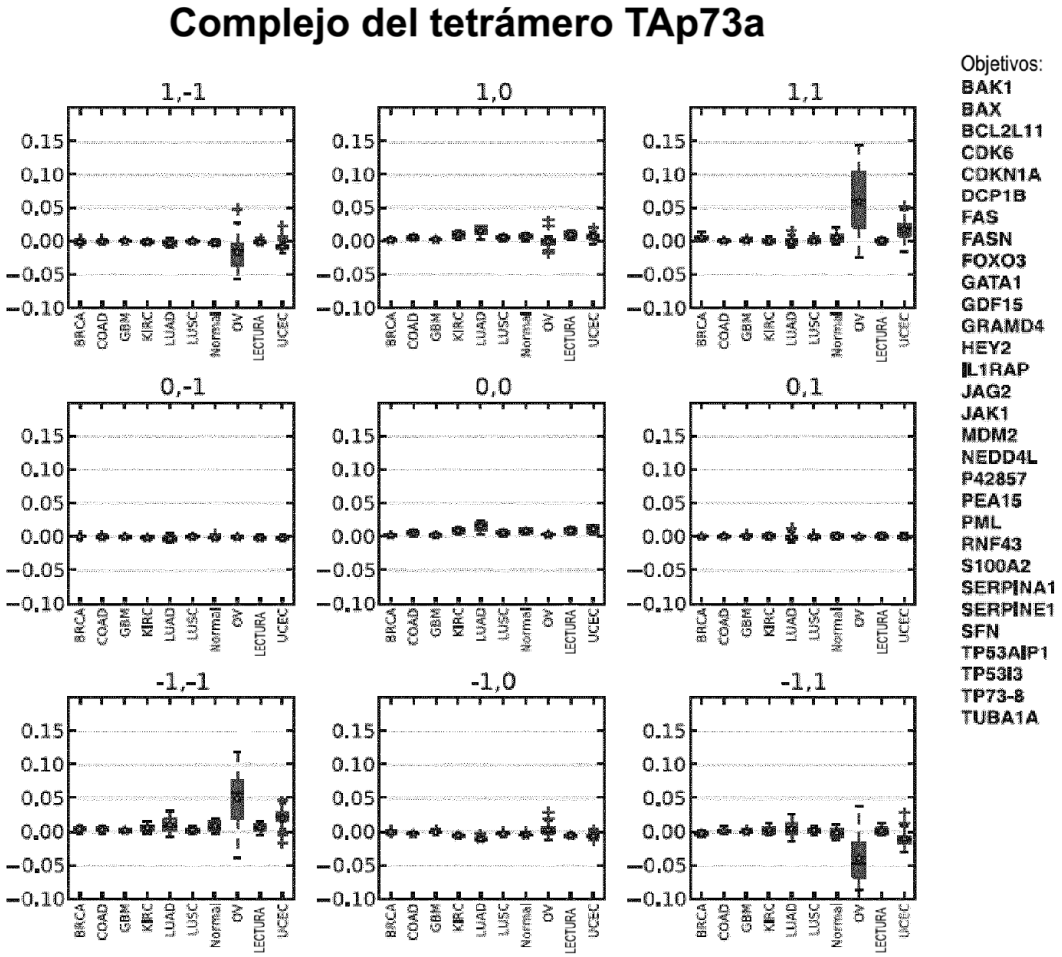

Complejo del tetrámero TAp73a
1,-1
1,0
1,1
0,-1
0,0
0,1
-1,-1
-1,0
-1,1
0.15
0.10
0.05
0.00
−0.05
−0.10
BRCA
COAD
GBM
KIRC
LUAD
LUSC
Normal
OV
LECTURA
UCEC
Objetivos:
BAK1
BAX
BCL2L11
CDK6
CDKN1A
DCP1B
FAS
FASN
FOXO3
GATA1
GDF15
GRAMD4
HEY2
IL1RAP
JAG2
JAK1
MDM2
NEDD4L
P42857
PEA15
PML
RNF43
S100A2
SERPINA1
SERPINE1
SFN
TP53AIP1
TP53I3
TP73-8
TUBA1A

Fig. 8B