(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 980 676**

(51) Int. Cl.:

G01N 15/14 (2014.01)
G06F 17/18 (2006.01)
G06V 10/75 (2012.01)
G06V 20/69 (2012.01)
G16B 40/30 (2009.01)
G06V 10/764 (2012.01)
G06F 18/213 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **23.05.2018 PCT/US2018/034199**

(87) Fecha y número de publicación internacional: **29.11.2018 WO18217933**

(96) Fecha de presentación y número de la solicitud europea: **23.05.2018 E 18731650 (0)**

(97) Fecha y número de publicación de la concesión europea: **14.02.2024 EP 3631417**

(54) Título: **Visualización, análisis comparativo y detección de diferencias automatizada para grandes conjuntos de datos multiparamétricos**

(30) Prioridad:

25.05.2017 US 201762511342 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
02.10.2024

(73) Titular/es:

FLOWJO, LLC (100.0%)
385 Williamson Way
Ashland, Oregon 97520, US

(72) Inventor/es:

ROEDERER, MARIO y
STADNISKY, MICHAEL, D.

(74) Agente/Representante:

PONTI & PARTNERS, S.L.P.

DESCRIPCIÓN

Visualización, análisis comparativo y detección de diferencias automatizada para grandes conjuntos de datos multiparamétricos

REFERENCIA CRUZADA A SOLICITUD(ES) RELACIONADA(S)

[0001] Esta solicitud reivindica el beneficio de prioridad, en virtud del artículo 119(e) del Título 35 del Código de Estados Unidos de América (*35 U.S.C. § 119(e)*), con respecto a la Solicitud de Patente Provisional de Estados Unidos de América n.º 62/511342, presentada el 25 de mayo de 2017 y que lleva por título *APPLIED COMPUTER TECHNOLOGY FOR VISUALIZATION, COMPARATIVE ANALYSIS, AND AUTOMATED DIFFERENCE DETECTION FOR LARGE MULTI-PARAMETER DATA SETS* ("TECNOLOGÍA INFORMÁTICA APLICADA PARA LA VISUALIZACIÓN, EL ANÁLISIS COMPARATIVO Y LA DETECCIÓN DE DIFERENCIAS AUTOMATIZADA PARA GRANDES CONJUNTOS DE DATOS MULTIPARAMÉTRICOS".

ANTECEDENTES

Campo técnico

[0002] Esta divulgación se refiere en general al campo de la evaluación automatizada de partículas, y más en concreto al análisis de muestras asistido por ordenador y a elementos de caracterización de partículas para grandes conjuntos de datos multiparamétricos.

Antecedentes

[0003] Los analizadores de partículas, como por ejemplo los citómetros de flujo y escaneado, son herramientas analíticas que permiten la caracterización de partículas basándose en medidas electro-ópticas como la dispersión de luz y la fluorescencia. En un citómetro de flujo, por ejemplo, partículas tales como moléculas, microesferas unidas por analitos o células individuales en una suspensión de fluido pasan a través de una región de detección en la que las partículas se exponen a una luz de excitación, normalmente procedente de uno o más láseres, y se miden las propiedades de dispersión de luz y fluorescencia de las partículas. Las partículas o componentes de las mismas normalmente se marcan con colorantes fluorescentes para facilitar la detección. Pueden detectarse simultáneamente una multiplicidad de diferentes partículas o componentes mediante el uso de colorantes fluorescentes espectralmente distintos para marcar las diferentes partículas o componentes. En algunas implementaciones se incluyen en el analizador una multiplicidad de fotodetectores, uno para cada uno de los parámetros de dispersión que se van a medir y uno o más para cada uno de los distintos colorantes que se van a detectar. Por ejemplo, algunas realizaciones incluyen configuraciones espectrales en las que se utilizan más de un sensor o detector por colorante. Los datos obtenidos comprenden las señales medidas para cada uno de los parámetros de dispersión de luz y las emisiones de fluorescencia.

[0004] Los analizadores de partículas pueden comprender además medios para registrar los datos medidos y analizar los datos. Por ejemplo, el almacenamiento y análisis de datos puede llevarse a cabo utilizando un ordenador conectado al equipo electrónico de detección. Por ejemplo, los datos se pueden almacenar en una forma tabular, donde cada fila corresponde a datos para una partícula y las columnas corresponden a cada una de las características medidas. El uso de formatos de archivo estándar, como por ejemplo el formato de archivo "FCS" (Estándar de Citometría de Flujo, por sus siglas en inglés, *Flow Cytometry Standard*), para almacenar datos procedentes de un analizador de partículas, facilita el análisis de datos utilizando programas y/o máquinas independientes. Mediante el uso de los métodos de análisis actuales, los datos normalmente se muestran en histogramas unidimensionales o gráficos bidimensionales (2D) para facilitar su visualización, pero se pueden utilizar otros métodos para visualizar los datos multidimensionales.

[0005] Los parámetros medidos usando, por ejemplo, un citómetro de flujo incluyen normalmente luz en la longitud de onda de excitación dispersada por la partícula en un ángulo estrecho a lo largo de una dirección principalmente hacia adelante, denominada dispersión frontal (FSC por sus siglas en inglés, *forward scatter*), la luz de excitación que es dispersada por la partícula en una dirección ortogonal al láser de excitación, denominada dispersión lateral (SSC por sus siglas en inglés, *side scatter*), y la luz emitida por moléculas fluorescentes en uno o más detectores que miden la señal en un rango de longitudes de onda espectrales, o por el colorante fluorescente que es detectado principalmente en ese detector o conjunto de detectores específicos. Se pueden identificar diferentes tipos de células por sus características de dispersión de luz y las emisiones de fluorescencia resultantes de marcar varias proteínas celulares u otros componentes con anticuerpos marcados con colorante fluorescente u otras sondas fluorescentes.

[0006] Tanto los citómetros de flujo como los de escaneado son comercializados, por ejemplo, por BD Biosciences (San José, California, Estados Unidos de América). La citometría de flujo se describe en, por ejemplo, Landy *et al.* (eds.), Clinical Flow Cytometry, Annals of the New York Academy of Sciences Volume 677 (1993); Bauer *et al.* (eds.), Clinical Flow Cytometry: Principles and Applications, Williams & Wilkins (1993); Ormerod (ed.), Flow Cytometry: A

Practical Approach, Oxford Univ. Press (1997); Jaroszeski *et al.* (eds.), Flow Cytometry Protocols, Methods in Molecular Biology n.º 91, Humana Press (1997) y Practical Shapiro, Flow Cytometry, 4.ª ed., Wiley-Liss (2003). La microscopía de imagen de fluorescencia se describe, por ejemplo, en Pawley (ed.), Handbook of Biological Confocal Microscopy, 2.ª edición, Plenum Press (1989).

[0007] Los datos obtenidos de un análisis de células (u otras partículas) mediante determinados analizadores de partículas, como por ejemplo una citometría de flujo de múltiples colores, son multidimensionales, en donde cada célula corresponde a un punto en un espacio multidimensional definido por los parámetros medidos. Las poblaciones de células o partículas se identifican como clústeres de puntos en el espacio de datos. La identificación de clústeres y, por consiguiente, de poblaciones, puede llevarse a cabo manualmente al dibujar una ventana (*gate*) alrededor de una población mostrada en uno o más gráficos bidimensionales, denominados "gráficos de dispersión" (*scatter plots*) o "gráficos de puntos" (*dot plots*), de los datos. Alternativamente, se pueden identificar los clústeres y se pueden determinar automáticamente las ventanas que definen los límites de las poblaciones. Se han descrito ejemplos de métodos para el acotamiento de zonas de interés (*gating*) automático en, por ejemplo, las patentes de Estados Unidos con números 4.845.653, 5.627.040, 5.739.000, 5.795.727, 5.962.238, 6.014.904, 6.944.338 y la publicación de patente de Estados Unidos n.º 2012/0245889.

[0008] En *Frequency Difference Gating: a multivariate method for identifying subsets that differ between samples* ("Acotamiento de zonas de interés por diferencia de frecuencia: un método multivariado para identificar subconjuntos que difieren entre muestras"), Mario Roederer *et al.*, se divulga una evaluación del acotamiento de zonas de interés por diferencia de frecuencia en varios escenarios de prueba para evaluar su utilidad. Se incluye la comparación de subconjuntos de células mononucleares de sangre periférica (PBMC por sus siglas en inglés, *peripheral blood mononuclear cell*), identificados únicamente mediante tinción de inmunofluorescencia: basándose en este conjunto de datos de aprendizaje, el algoritmo generó automáticamente una ventana de dispersión frontal y lateral precisa para identificar linfocitos. Además, se aplicó el algoritmo para identificar diferencias sutiles entre los subconjuntos de memoria CD4 basándose en datos de inmunofenotipado de 8 colores. La ventana tridimensional resultante podría resolver subconjuntos de células mucho más frecuentes en un subconjunto que en el otro; ninguna combinación de ventanas bidimensionales podía lograr esta resolución. Por último, se utilizó el algoritmo para comparar poblaciones de células B derivadas de ratones de diferentes edades o cepas, y se descubrió que el algoritmo podía encontrar diferencias muy sutiles entre las poblaciones.

[0009] En *Mapping cell populations in flow cytometry data for cross-sample comparison using Friedman-Rafsky test statistic as a distance measure* ("Asignación de poblaciones celulares en datos de citometría de flujo para la comparación entre muestras utilizando estadística de la prueba de contraste de Friedman-Rafsky como medida de distancia"), Chiaowen Hsiao *et al.*, se divulga FlowMap-FR, un método novedoso para la asignación de poblaciones celulares en muestras de citometría de flujo. FlowMap-FR se basa en la prueba estadística de contraste no paramétrica de Friedman-Rafsky (estadística FR), la cual cuantifica la equivalencia de distribuciones multivariadas. Tal y como se aplica a los datos de citometría de flujo mediante FlowMap-FR, el estadístico FR cuantifica objetivamente la similitud entre poblaciones celulares en función de las formas, tamaños y posiciones de las distribuciones de datos de fluorescencia en el espacio de características multidimensionales. Para probar y evaluar el rendimiento de FlowMap-FR, se simularon los tipos de variaciones de muestras biológicas y técnicas que se observan comúnmente en los datos de citometría de flujo. Los resultados muestran que FlowMap-FR es capaz de identificar eficazmente poblaciones de células equivalentes entre muestras en escenarios de diferencias de proporción y cambios de posición modestos.

RESUMEN

[0010] Cada uno de los sistemas, métodos y dispositivos de la divulgación tiene varios aspectos innovadores, ninguno de los cuales es el único responsable de los atributos deseables divulgados en el presente.

[0011] De conformidad con la invención, se proporciona un método implementado por ordenador para visualizar diferencias entre conjuntos de datos de n dimensiones de acuerdo con la reivindicación 1. El método implementado por ordenador se lleva a cabo bajo el control de uno o más dispositivos de procesamiento.

[0012] En algunas implementaciones del método implementado por ordenador, el primer y segundo conjunto de datos incluyen datos de muestras de células multiparamétricas.

[0013] En algunas implementaciones, el método implementado por ordenador también incluye ajustar el umbral definido en respuesta a una entrada del usuario y ajustar la visualización en función del umbral definido ajustado. El umbral definido puede incluir o representar una pluralidad de umbrales definidos. El método implementado por ordenador puede incluir la generación de la visualización mediante la codificación por colores de las regiones basándose al menos en parte en el acotamiento de zonas de interés por diferencia de frecuencia. El umbral definido puede incluir un umbral superior que identifica una o más regiones clasificadas como que tienen una mayor frecuencia de eventos del primer subconjunto que del segundo subconjunto. Adicional o alternativamente, el umbral definido puede incluir un umbral inferior que identifica una o más regiones que tienen una mayor frecuencia de eventos del segundo subconjunto que del primer subconjunto. En algunas implementaciones, el umbral definido incluye límites de

rango medio que identifican una o más regiones que tienen una frecuencia de eventos similar entre el primer subconjunto y el segundo subconjunto.

[0014] Cuando se utilizan histogramas binormalizados, la generación de la visualización puede incluir la representación de un mapa de calor de los histogramas de diferencia binormalizados.

[0015] Algunos ejemplos del método incluyen la generación de un tercer conjunto de datos basado en al menos una ventana definida por un usuario a través de la visualización. El primer conjunto de datos puede incluir una muestra de control, como por ejemplo datos celulares de materia sana o datos celulares de materia cancerosa.

[0016] En otro aspecto innovador, se da a conocer un sistema de acuerdo con la reivindicación 14.

[0017] En algunas implementaciones, el umbral incluye al menos uno de los siguientes elementos: un umbral superior que identifica una o más regiones clasificadas como que tienen una mayor frecuencia de eventos del primer subconjunto que del segundo subconjunto, un umbral inferior que identifica una o más regiones que tienen una mayor frecuencia de eventos del segundo subconjunto que del primer subconjunto, o límites de rango medio que identifican una o más regiones que tienen una frecuencia de eventos similar entre el primer subconjunto y el segundo subconjunto.

[0018] Los elementos del histograma multidimensional pueden ser intervalos (*bins*).

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0019]

La Figura 1 ilustra un ejemplo de sistema informático que se puede utilizar para dar soporte a las técnicas innovadoras de procesamiento y visualización de datos descritas en el presente.
La Figura 2A muestra ejemplos de conjuntos de datos de expresión génica celular.
La Figura 2B muestra una vista de tabla de ejemplos de datos de expresión génica celular.
La Figura 3 muestra un ejemplo de flujo de proceso para un método de acotamiento de zonas de interés y visualización por diferencia de frecuencia.
Las Figuras 4A-4C muestran ejemplos de visualizaciones de acotamiento de zonas de interés por diferencia de frecuencia que pueden generarse.

DESCRIPCIÓN DETALLADA

[0020] Una sola célula puede representar la unidad básica de la enfermedad, pero las tecnologías emergentes en análisis de citometría de flujo (>40 parámetros por célula) y secuenciación de células individuales (entre 10 000 y más de 60 000 parámetros por célula) pueden verse obstaculizadas por pasos manuales miopes, secuenciales y que requieren mucho tiempo o enfoques de reducción de datos no deterministas y costosos desde un punto de vista informático. Este es un problema bien documentado en la técnica, pero la técnica se ha encontrado con dificultades para desarrollar soluciones a este problema. Esta falta de determinismo impide una comparación significativa de muestras que sustente todos los tipos de análisis de datos, pero especialmente las ciencias biológicas: la comparación con una medida de control o “normal” saludable es un componente crítico, pero la ciencia de células individuales emplea su mejor capacidad para enfrentarse a estas grandes dificultades y realizar estas comparaciones significativas en un espacio fenotípico muy ampliado.

[0021] Se describen características para abordar los problemas presentados por grandes conjuntos de datos multiparamétricos, los cuales no pueden ser explorados fundamentalmente por un ser humano ni pueden ser sometidos a comparaciones significativas. Especialmente cuando se trata de comparaciones dentro de la misma muestra o entre diferentes muestras, los seres humanos (por ejemplo, los expertos científicos) pueden mostrar un sesgo increíble causado por sus conocimientos y experiencia previos. Se ha demostrado que las principales diferencias que impulsan las diferencias biológicas pueden en realidad deberse a subconjuntos de células que no se incluyen en el análisis manual de datos mediante el acotamiento manual de zonas de interés (véase M. D. Stadnisky, S. Siddiq, J. Almarode, J. Quinn, A. Hart, *Reproducible Reduction: Deterministic tSNE using regression trees enables intra-sample comparison* (“Reducción reproducible: tSNE determinista que utiliza árboles de regresión permite la comparación dentro de muestras”) CYTO 2016: XXXI Congreso de la Sociedad Internacional para el Avance de la Citometría. Seattle, Washington. Junio de 2016). Es decir, los investigadores en un estudio pueden pasar por alto por completo los fenotipos responsables de una respuesta biológica o la diferencia entre pacientes sanos y el estado de la enfermedad. Más adelante se analizan en mayor detalle dos ejemplos para poner de manifiesto la limitación de estos enfoques en la comparación de datos de partículas.

[0022] Para usar un ejemplo específico, procedente de un estudio que examina cómo el cuerpo cuenta y regula el número de células inmunitarias (véase Roederer *et. al. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis* (“La arquitectura genética del sistema inmunológico humano: un biorrecurso para la autoinmunidad y la patogénesis de enfermedades”). Cell. Abril de 2015 9;161(2):387-403. doi:

10.1016/j.cell.2015.02.046. Epub 2015 Mar 12.), existen muchas poblaciones posibles ("rasgos"; véase *Another Application of Technology* ("Otra aplicación de la tecnología") que se conocen:

Canónicos: subconjuntos predefinidos, "conocidos" o descritos.

$$T_{CM} = CD45RA^{-}CCR7^{+}CD28^{+}$$

$$T_{REG} = CD45RO^{+}CD127^{-}CD25^{+}CD39^{+}$$

[0023] Para un panel determinado de anticuerpos, es posible que se puedan identificar más de 50 poblaciones canónicas. Sin embargo, y como se mostrará en dos ejemplos específicos más adelante, este enfoque omite muchos subconjuntos. Además, el enfoque se basa en el supuesto fundamental de que las poblaciones canónicas son: (1) adecuadamente definidas; y (2) bien conocidas. Sin embargo, ¿qué ocurre si las poblaciones canónicas no han sido definidas adecuadamente? El análisis de secuenciación de células individuales muestra, de manera imparcial, lagunas analíticas en el uso de paneles canónicos actuales para dividir en subconjuntos e identificar células. Por ejemplo, en un estudio reciente realizado en células linfoides innatas (CLI), definimos 3-5 marcadores adicionales no declarados previamente para cada subconjunto canónico de CLI y definimos 3 subconjuntos secundarios "nuevos" para cada subconjunto canónico. De hecho, cada célula es única por definición, por lo que un factor clave es analizar lo que sabemos de los subconjuntos canónicos y combinarlo con el enfoque innovador que se describe en el presente.

[0024] Un proceso analítico alternativo puede incluir analizar todo a la vez, es decir, todas las combinaciones posibles de marcadores. Una ventaja de este enfoque es que el análisis no pasará por alto ninguna combinación y todavía se pueden identificar poblaciones canónicas. Sin embargo, una desventaja de este enfoque es que, dado que requiere ejecutar algoritmos que computan sobre el conjunto de datos de n dimensiones, se requiere una gran cantidad de recursos. Como ejemplo de la cantidad de datos, hemos ejecutado estos algoritmos en conjuntos de datos que examinan 12 proteínas y más de 60 000 ARNm y mutaciones de splicing en cada célula para miles de células. Esto puede considerarse un experimento de bajo rendimiento, en comparación con otros experimentos que analizan más de 100 000 parámetros por célula individual para muchas muestras, cada una de las cuales tiene cientos de miles de células. Fundamentalmente, este tipo de rendimiento se ve obstaculizado por la incapacidad de comparar los tratamientos y el estado de la enfermedad, por lo que muchos estudios basados en análisis siguen siendo de naturaleza descriptiva. En otro ejemplo, procedente de un estudio que examina cómo el cuerpo cuenta y regula el número de células inmunes usando una sola modalidad (por ejemplo, siete paneles examinados mediante citometría de flujo), había:

59 "linajes".
77 941 subconjuntos totales.
684 valores de MFI (intensidad de fluorescencia media).

Total: 78 683 rasgos

... y esto es antes de combinarlos con otros flujos de datos disponibles para cada par gemelo estudiado.

[0025] Los recursos necesarios para procesar dichos conjuntos de datos multidimensionales de gran tamaño pueden exceder los recursos disponibles o esperados para proporcionar un resultado en un período de tiempo práctico. Los recursos pueden incluir recursos informáticos, recursos de energía, recursos de memoria, recursos de red, recursos de transceptor o similares.

[0026] Con el fin de conseguir que los descubrimientos en un espacio de n dimensiones sean más faciles de gestionar, la reducción de datos puede ser una técnica de visualización útil para reducir el número de variables aleatorias sometidas a consideración y obtener un conjunto de variables principales "no correlacionadas". Esto proporciona un método visual para explorar nuevos parámetros que representan una proyección de los datos de n dimensiones, que a su vez se puede analizar más a fondo. No obstante, existen dos enfoques para la reducción de datos: la extracción de características y la selección de características; y la extracción de características se ha utilizado ampliamente en la ciencia de células individuales, particularmente en el análisis de componentes principales (ACP) y la incrustación de vecinos estocásticos distribuidos en t (t-SNE, *t-distributed stochastic neighbor embedding*).

[0027] Se puede utilizar el ACP para extraer variables linealmente no correlacionadas que se denominan componentes principales (por ejemplo, nuevos parámetros) y que representan la varianza en los datos subyacentes (por ejemplo, datos "sin procesar"). Sin embargo, nosotros y otros hemos demostrado que la mayor ventaja del ACP representa también su mayor desventaja: "El ACP encuentra que la representación más óptima dentro del conjunto de posibles proyecciones lineales de los datos también constituye una limitación importante: una proyección lineal puede ser demasiado restrictiva para producir representaciones precisas". Shekhar, Karthik *et al.*, *Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE)* ("Clasificación automática de la expresión celular mediante incrustación estocástica no lineal (ACCENSE))". *Proceedings of the National Academy of Sciences* (Actas de la Academia Nacional de Ciencias de los Estados Unidos de América) 111.1 (2014): 202-207. Según nuestros

descubrimientos, el ACP es incapaz de identificar clústeres identificados por un experto científico en un conjunto de datos de baja dimensión (por ejemplo, ocho parámetros). Además, en la secuenciación de células individuales, el ACP es sensible al número de transcripción.

[0028] A fin de superar estas limitaciones, se han realizado muchos trabajos recientes en la aplicación de t-SNE a datos de células individuales. La t-SNE es una poderosa técnica de extracción de características no lineales. Los aspectos de t-SNE se describen en Van der Maaten, Laurens y Geoffrey Hinton. *Visualizing data using t-SNE* ("Visualización de datos utilizando t-SNE"). Journal of Machine Learning Research 9.2579-2605 (2008): 85 y Van Der Maaten, Laurens, Eric Postma y Jaap Van den Herik. *Dimensionality reduction: a comparative* ("Reducción de dimensionalidad: una comparativa"). Journal of Machine Learning Research 10 (2009): 66-71. La t-SNE modela cada objeto de alta dimensión mediante un punto bidimensional o tridimensional, de tal manera que los objetos similares se modelan mediante puntos cercanos y los objetos diferentes se modelan mediante puntos distantes. Esto resulta útil en la visualización y el análisis biológicos, ya que los vecinos más cercanos reflejan células similares que pueden agruparse en un subconjunto.

[0029] La t-SNE se ha adaptado para citometría de células individuales en dos enfoques diferentes usando la adición de (1) partición y rendimiento, así como agrupación (Amir, El-ad David *et al.*, *viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia* ("viSNE permite la visualización de datos de células individuales de altas dimensiones y revela la heterogeneidad fenotípica de la leucemia". Nature biotechnology 31.6 (2013): 545-552) y (2) agrupación y una aplicación (Shekhar, Karthik *et al. Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE)* ("Clasificación automática de la expresión celular mediante incrustación estocástica no lineal (ACCENSE)". *Proceedings of the National Academy of Sciences* (Actas de la Academia Nacional de Ciencias de los Estados Unidos de América) 111.1 (2014): 202-207). La t-SNE ha demostrado ser una técnica prometedora y se ha utilizado con éxito para identificar poblaciones en la secuenciación de células individuales (Macosko, Evan Z *et al.*, *Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets* ("Elaboración de perfiles de expresión de todo el genoma altamente paralelos de células individuales utilizando gotitas de nanolitros"). Cell 161.5 (2015): 1202-1214.; Tirosh, Itay *et al.*, *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* ("Disección del ecosistema multicelular del melanoma metastásico mediante secuenciación de ARN de células individuales"). Science 352.6282 (2016): 189-196).

[0030] Sin embargo, aún quedan algunos desafíos interrelacionados críticos. Un desafío es el gasto de recursos informáticos. Para los sistemas analíticos que incluyen t-SNE, el gasto de recursos informáticos se cuantifica como un tiempo de ejecución lento que se escala de forma deficiente, $O(N^2)$ para la implementación de referencia u O(N log N) para la implementación Barnes-Hut. El gasto de recursos informáticos depende del número de parámetros/rasgos, pero también del número de mediciones de partículas independientes (por ejemplo, eventos, células, etc.). La ejecución de miles de parámetros y eventos puede tener como consecuencia un tiempo de ejecución prolongado mientras se espera el análisis, o simplemente no es factible sin grandes recursos informáticos, como por ejemplo un grupo de superordenador/servidor.

[0031] Otro desafío está relacionado con el determinismo que se ha mencionado anteriormente. La t-SNE aprende una asignación no paramétrica de los datos y, por lo tanto, no existe un método fácilmente obtenible para la estimación fuera de muestra. El resultado es que dos ejecuciones de t-SNE en el mismo conjunto de datos probablemente produzcan dos visualizaciones diferentes. Esta ausencia de determinismo es inherente al algoritmo y requiere la concatenación de archivos de datos, por ejemplo, combinando tejidos sanos y enfermos en un solo archivo y luego ejecutando t-SNE en el archivo de datos combinado para intentar visualizar las diferencias. Esto se explica en mayor detalle en el Ejemplo n.º 2 que se mostrará más adelante.

[0032] Otro desafío está relacionado con las comparaciones analíticas que se pueden realizar. La comparación exacerba en extremo los dos desafíos antes mencionados: realizar una comparación directa por pares de N muestras lleva a comparaciones (N(N-1))/2 (por lo tanto, el rendimiento de N^2), pero la comparación de un número N de muestras es peor que este rendimiento informático ya deficiente. Frente a este desafío, puede ocurrir cierta preselección de parámetros/rasgos, puesto que no todos los rasgos son útiles, ya que: (1) algunos se consideran "triviales" (células demasiado infrecuentes para merecer un análisis más detallado (<0,5% de un linaje)); y (2) algunos son demasiado variables (ya sea en el ensayo o biológicamente) basándose en los controles de variabilidad longitudinal e intraensayo. Sin embargo, para fines analíticos, no se tiene ninguna orientación sobre dónde enfocarse en el espacio de n dimensiones. Sin un medio para realizar comparaciones significativas, una respuesta es centrarse en diferencias obvias, que pueden no reflejar con precisión la realidad y también pueden ser difíciles de identificar de manera analíticamente rigurosa.

[0033] Las características descritas en el presente proporcionan un método estadísticamente riguroso para comparar conjuntos de datos complejos y de alta dimensión en cualquier subgrupo de sujetos y proporciona, por primera vez, un enfoque de fenotipado profundo de diagnóstico para datos de múltiples parámetros.

[0034] El contenido de ensayos de células individuales ha aumentado sustancialmente en los últimos años, pero la capacidad de realizar la tarea más fundamental (la comparación) utilizando estos conjuntos de datos de múltiples

parámetros está muy limitada. En términos de órdenes de magnitud, el campo ha pasado de observar la biología con binoculares gigantes (por ejemplo, con un aumento de 40X) a usar el telescopio Hubble (por ejemplo, con un aumento de 8000X), pero el campo carece de una forma significativa de comparar las diferencias entre los cuerpos celestes que observamos. Existen sistemas y métodos para ver y descubrir, pero estos no pueden realizar un trabajo analítico significativo.

[0035] Las características descritas, cuyos ejemplos de realizaciones se divulgan en el presente, superan las limitaciones de la biología de células individuales. Las características desbloquean datos de n dimensiones para utilizarlos como herramienta de diagnóstico. Se puede utilizar el enfoque descrito para estratificar a los pacientes para la medicina de precisión, reducir el tiempo de análisis manual, proporcionar un mecanismo que revele de forma rápida y reproducible poblaciones de células no identificadas y crear la oportunidad de comparar estudios de gran tamaño de células individuales.

[0036] Por definición, un inmunólogo no es un biólogo de sistemas y, como muestra el ejemplo anterior que utiliza poblaciones "canónicas", este científico puede especializarse en estudiar en profundidad un grupo de tipos de células disponibles cuando intenta comprender una enfermedad o una función de célula inmunitaria. Esto se debe a que convertirse en un experto en muchos tipos de células –su funcionalidad, interacciones y especialmente identificación – requiere mucho tiempo y formación.

[0037] Consideremos dos ejemplos reales que ilustran este sesgo de poblaciones conocidas/canónicas y muestran la necesidad evidente de una comparación rigurosa como facilitador de las ciencias biológicas y la medicina de precisión. Estos ejemplos deberán entenderse como ilustrativos de problemas más amplios en las ciencias biológicas.

Ejemplo 1: linfocitos T CD8+ en un estudio de vacunas

[0038] En un estudio de citometría de múltiples parámetros de linfocitos T CD8+ después de la vacunación, los autores limitaron sus análisis y resultados a los subconjuntos de linfocitos T CD8+ canónicas (Newell *et. al.*, *Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8+ T Cell Phenotypes Immunity* ("La citometría por tiempo de vuelo muestra la expresión combinatoria de citocinas y nichos celulares específicos de virus dentro de un continuo de inmunidad de los fenotipos de linfocitos T CD8+"), 2012). Analizamos estos datos más a fondo, comenzando por concatenar (por ejemplo, combinar) los datos de linfocitos T CD8+ de 6 pacientes diferentes. Después redujimos el espacio de datos de 25 parámetros usando t-SNE.

[0039] Los datos de linfocitos T CD8+ concatenados para los 6 pacientes diferentes incluyen claramente dos firmas inmunes diferentes. Nuestro experimento descubrió que 4 de los 6 pacientes se distinguían por tipos de células que quedaban fuera de las ventanas tradicionales (por ejemplo, poblaciones "canónicas").

[0040] Esta diferencia en las firmas inmunes ilustra las limitaciones de los enfoques analíticos actuales. En las implementaciones de la medicina de precisión, los marcos actuales no proporcionan una forma rigurosa de comparar la firma inmune específica de los pacientes (por ejemplo, respuesta de los linfocitos T CD8+). Ello sugiere que identificar diferencias entre pacientes puede resultar casi imposible si se utilizan las técnicas analíticas actuales. Esto, a su vez, limita la capacidad de identificar la totalidad de una firma inmunitaria y, por tanto, dificulta la práctica de la medicina de precisión.

[0041] En las implementaciones de descubrimiento, los marcos existentes no proporcionan una forma rigurosa de extraer todos los subconjuntos de células que pueden diferenciar dos respuestas completamente diferentes a la vacuna. La posibilidad de que existan tipos de células fuera del canon ilustra cómo la concentración en las poblaciones canónicas puede influir en un estudio publicado. Esto también pone de manifiesto que no existe una herramienta para dividir simplemente los dos grupos de pacientes e "interrogar a los datos" para descubrir la diferencia.

Ejemplo 2: Las células linfoides innatas en la respuesta inmunitaria frente al cáncer y las respuestas inmunitarias específicas de los tejidos

[0042] En otro ejemplo de citometría de múltiples parámetros, realizamos un metaanálisis que compara las respuestas inmunes en tejido sano y tumor en tres órganos diferentes: colon, hígado y pulmones.

[0043] Cabe destacar que hemos observado el mismo problema de comparación en la secuenciación de células individuales, donde nos gustaría extraer rigurosamente las células que son "plásticas" o quizás se están diferenciando en otro subconjunto de células, representadas por los puntos de datos de un color determinado que parece ser parte de otro subconjunto de células.

[0044] Por ejemplo, consideremos un gráfico de dispersión de t-SNE de 847 de los genes expresados de forma más diferencial en subconjuntos de células inmunitarias (de 60 000 ARNm y mutaciones de splicing). El color se puede utilizar en el diagrama de dispersión para representar el fenotipo utilizando definiciones "canónicas" mediante citometría de flujo. Los 847 genes pueden incluir células cuya inclusión en clústeres no se "predice" por las definiciones

de flujo canónicas. Sin embargo, debido a que el gráfico de dispersión presenta los genes expresados diferenciados solo por el color, actualmente no es posible extraer estas células que son diferentes de manera rigurosa y automática. Estas células pueden estar ocultas entre las células analizadas canónicamente, lo que hace que no se detecten diferencias potencialmente significativas.

[0045] Por lo tanto, como ilustran los ejemplos anteriores, en lugar de la maldición de la dimensionalidad [R. E. Bellman; Corporación Rand (1957). *Dynamic Programming* ("Programación dinámica"). Princeton University Press. Republicado: Richard Ernest Bellman (2003). *Dynamic Programming* ("Programación dinámica"). Courier Dover Publications, y Richard Ernest Bellman (1961). *Adaptive Control Processes: a guided tour* ("Procesos de control adaptativo: una visita guiada"). Princeton University Press.], la ventana de fenotipo ampliada, en la que cada subconjunto de células tiene un significado biológico y podría correlacionarse con una enfermedad, pero que puede ser pasada por alto por el experto, presenta una discordancia de conocimientos que tiene como consecuencia una enfoque centrado en los fenotipos conocidos y que se preste poca atención a otros subconjuntos de células. Esto tiene como resultado que grandes cantidades de datos permanezcan latentes en el proceso de descubrimiento en todos los niveles de citometría; por ejemplo, un ensayo "estándar" de 10 colores = 1024 posibles fenotipos de interés. Además, sin ninguna forma de poder realizar comparaciones entre muestras, ¿cómo se supone que un biólogo puede centrar su atención dentro del espacio de datos, por ejemplo, en los fenotipos que realmente se correlacionan con la enfermedad y, por lo tanto, diferencian o impulsan dicha enfermedad?

[0046] No obstante, la recopilación de más parámetros y la participación en este proceso de descubrimiento no es un ejercicio de "más es más", sino que resulta fundamental para encontrar correlaciones de morbilidad o eficacia terapéutica. Supongamos que un investigador está buscando un subconjunto de células T definido por una combinación de 4 parámetros (también conocidos como "marcadores") que es importante en una respuesta inmune específica. Cuando se utilizan menos de cuatro marcadores, el investigador incluye otras poblaciones de células irrelevantes en nuestro análisis, lo que diluye su capacidad para detectar las células de interés. A medida que se utilizan cada vez menos marcadores, cada vez se miden más células irrelevantes, lo que aumenta el "ruido" y, en consecuencia, reduce la detección de las células importantes, es decir, aquellas que se correlacionan con la protección. En general, es más difícil encontrar asociaciones significativas cuando se realizan mediciones masivas. Sin embargo, *a priori* se desconoce el número de marcadores necesarios para encontrar una correlación de protección. Es casi seguro que las respuestas protectoras comprenden células que expresan un patrón de funciones múltiples. Por tanto, al examinar más marcadores en más células, el sistema puede identificar subconjuntos de células que pueden correlacionarse con la morbilidad o la eficacia terapéutica. Utilizando las características innovadoras descritas, se pueden identificar subconjuntos nuevos e inesperados de importancia en las enfermedades.

[0047] Por consiguiente, los científicos que intentan servirse de las tecnologías de células individuales en la investigación centrada en el descubrimiento más allá de un área de enfoque estrecha se enfrentan a una tarea difícil, no determinista y no reproducible.

[0048] Existen algunas soluciones convencionales en la técnica que podrían emplearse para ejecutar análisis de descubrimiento en este conjunto de datos. Una es el análisis manual. El análisis manual puede incluir la revisión de gráficos de visualización de datos de eventos. Otra solución puede incluir estadísticas básicas como KS, análisis de chi-cuadrado de Cox. Pero esto puede resultar demasiado delicado. Además, las estadísticas no proporcionan un método de acotamiento de zonas de interés y normalmente se limitan al análisis univariado.

[0049] Otra solución puede incluir recurrir a un bioinformático. En un caso raro, un investigador colaborará con un bioinformático que pueda aprovechar sus competencias especializadas para analizar los datos.

[0050] Otro enfoque es utilizar la reducción para ayudar a enfocar el análisis de los datos multidimensionales. Un ejemplo de reducción es la visualización de árboles (SPADE, X-shift, flowSOM) o la inferencia de progresión (Wanderlust, Pseudotime). Sin embargo, estas reducciones no son deterministas y no proporcionan ningún elemento de comparación. Como se ha estudiado anteriormente, el ACP es otra opción de reducción, pero conlleva problemas de validación. La t-SNE es otra opción, pero ya se han descrito lo problemas de este enfoque anteriormente. Por ejemplo, en el procesamiento de reducción de datos t-SNE, la información de los datos sin procesar se pierde, pero el procesamiento intenta conservar la mayor "relación" posible. A pesar de estos esfuerzos, la t-SNE puede agrupar datos en 2 dimensiones de una manera que preserve la relación en regiones locales, pero no globales, dejando abierta la posibilidad de que se ignoren diferencias potencialmente significativas.

[0051] Los enfoques convencionales pueden ser problemáticos en varios aspectos. Existe una necesidad en la técnica de mejoras técnicas con respecto a cómo se puede aplicar la tecnología informática para identificar y visualizar de manera significativa diferencias destacadas entre muestras en grandes conjuntos de datos multiparamétricos. Como solución a este problema, se describen características para visualizaciones basadas en acotamiento de zonas de interés por diferencia de frecuencia (FDG).

[0052] La Figura 1 ilustra un ejemplo de sistema informático 100 que se puede utilizar para apoyar las técnicas innovadoras de procesamiento y visualización de datos descritas en el presente. El ejemplo de sistema informático 100

comprende un procesador 102, una memoria 104, una base de datos 106 y una pantalla 108 que pueden estar en comunicación entre sí a través de una tecnología de interconexión, como por ejemplo el bus 110.

[0053] El procesador 102 puede adoptar la forma de cualquier procesador apropiado para realizar las operaciones descritas en el presente. Por ejemplo, la CPU de un ordenador portátil o estación de trabajo sería adecuada para su uso como procesador 102. Se entenderá que el procesador 102 puede comprender múltiples procesadores, incluidos procesadores distribuidos que se comunican entre sí a través de una red para llevar a cabo tareas descritas en el presente (por ejemplo, recursos de procesamiento de informática en la nube). La memoria 104 puede adoptar la forma de cualquier memoria de ordenador adecuada para cooperar con el procesador 102 en la ejecución de las tareas descritas en el presente. Se entenderá que la memoria 104 puede adoptar la forma de múltiples dispositivos de memoria, incluida la memoria que se distribuye a través de una red. De forma similar, la base de datos 106 puede adoptar la forma de cualquier depósito de datos accesible al procesador 102 (por ejemplo, un sistema de archivos en un ordenador, una base de datos relacional, etc.), y deberá entenderse que la base de datos 106 puede adoptar la forma de múltiples bases de datos distribuidas (por ejemplo, almacenamiento en la nube). La pantalla 108 puede adoptar la forma de un monitor o pantalla de ordenador que sea capaz de generar las visualizaciones descritas en el presente.

[0054] Las características descritas en el presente son aplicables a conjuntos de datos de n dimensiones, que pueden adoptar la forma de datos de muestra 112 (por ejemplo, datos de expresión génica celular u otros datos de medición de partículas). Se pueden generar los datos de expresión génica celular mediante secuenciación de próxima generación (por ejemplo, para la medición de secuenciación de ARN (RNASeq) y secuenciación de ARN de células individuales (scRNA-Seq), entre otros enfoques de secuenciación). Sin embargo, esto es solo un ejemplo y pueden emplearse otras técnicas para generar datos de expresión génica celular. Ejemplos adicionales incluyen enfoques de reacción en cadena de la polimerasa que incluyen gotitas digitales y transcriptasa inversa. Otros ejemplos adicionales incluyen la medición de ARN mediante citometría de flujo y *microarrays*, entre otros, que producen archivos de datos que contienen la cuantificación de ADN y/o ARN, o mediante programas de software que procesan los datos leídos sin procesar (análisis primario y secundario) para generar los archivos de datos de expresión génica u otros marcadores biológicos.

[0055] Los datos de muestra 112 se pueden caracterizar como un gran conjunto de datos multiparamétricos que plantea desafíos técnicos especiales por lo que respecta a la dificultad para crear visualizaciones significativas, particularmente cuando se considera con respecto a la biología subyacente, de modo que la información biológicamente relevante se presente significativamente a los usuarios en una manera visual. Por ejemplo, los datos de expresión génica celular pueden comprender datos para un gran número de células individuales y poblaciones de células, con parámetros para cada célula o población de células que pueden extenderse a 10 000-60 000 o más parámetros. Los datos de muestra 112 pueden leerse de archivos en la base de datos 106 y cargarse en la memoria 104 como una pluralidad de estructuras de datos 116 para ser manipuladas por el procesador 102 durante la ejecución de un programa de análisis y visualización 114. El programa de análisis y visualización 114 puede comprender código informático ejecutable por procesador en forma de una pluralidad de instrucciones ejecutables por procesador que residen en un medio de almacenamiento no transitorio legible por ordenador, por ejemplo, la memoria 104.

[0056] La Figura 2A muestra ejemplos de conjuntos de datos de expresión génica celular en los que cada célula (o población celular) se identifica mediante un identificador (Id.) de célula (*Cell ID*) asociado con una pluralidad de parámetros, teniendo cada parámetro un identificador y un valor en relación con un identificador de célula. Como se ha indicado, los datos de expresión génica para las células son altamente dimensionales y el número de parámetros para cada célula puede alcanzar 10 000-60 000 o más parámetros, ya sea por célula o por población de células. Entre los ejemplos de parámetros en los datos celulares figuran recuentos de expresión génica en la célula en cuestión para una gran cantidad de genes. Por lo tanto, el parámetro 1 para la célula 1 puede corresponder al gen 1 y su valor puede ser un recuento de expresiones para el gen 1 en la célula 1. De manera similar, el parámetro 2 para la célula 1 puede corresponder al gen 2 y su valor puede ser un recuento de expresiones para el gen 2 en la célula 1.

[0057] La Figura 2B muestra una vista de tabla de ejemplos de datos de expresión génica celular. Cada fila de la tabla 200 corresponde a una célula diferente (véase la columna Célula), y las diversas columnas etiquetadas como Gen 1, Gen 2, etc. corresponden a diferentes genes y las células de la tabla identifican recuentos de expresiones génicas para los genes correspondientes en cada célula específica. Esta tabla también puede incluir parámetros distintos de los genes. Por ejemplo, los datos de expresión génica celular 112 pueden incluir valores de datos para parámetros tales como incrustación de vecinos estocástica distribuida en t (tSNE), el análisis de componentes principales (ACP), el análisis discriminante lineal (ADL), etc. en cada célula de la tabla, donde estos valores de datos representan un cálculo de análisis cuyo valor captura las diferencias para células individuales en n parámetros. Los datos de expresión génica celular 112 se pueden almacenar en cualquiera de una serie de formatos (por ejemplo, como archivos CSV, tablas de bases de datos (por ejemplo, como datos relacionales en una base de datos relacional), representaciones de datos dispersos, formatos binarios y otros).

[0058] La Figura 3 muestra un ejemplo de proceso de flujo para un método de acotamiento de zonas de interés por diferencia de frecuencia y visualización. El método puede implementarse total o parcialmente en uno o más de los dispositivos descritos. En algunas implementaciones, el programa de análisis y visualización 114 puede incluir instrucciones para implementar al menos parte del método mostrado. Los bloques 300-306 describen opciones para preparar datos de muestra para su análisis.

[0059] Según una primera opción, el sistema puede recibir archivos que se van a concatenar en el bloque 300 para la comparación entre diferentes muestras. Por ejemplo, un primer archivo puede corresponder a una muestra de prueba y un segundo archivo puede corresponder a una muestra de control. Cada archivo corresponde a datos de muestra multidimensionales, como los datos celulares que se muestran en las Figuras 2A y 2B. Recibir los archivos puede incluir recibir un archivo cargado desde el ordenador de un investigador. En algunas implementaciones, los archivos pueden recibirse desde un analizador de partículas, por ejemplo, un citómetro de flujo. Para mayor claridad, la discusión se refiere a dos archivos, pero la concatenación puede basarse en más de dos archivos.

[0060] En el bloque 300, el sistema concatena los dos archivos. La concatenación de los archivos puede incluir la generación de un nuevo parámetro en una tabla u otra estructura de datos que incluya los datos de muestra de los archivos para indicar el origen de una entrada. Por ejemplo, si se utiliza una tabla para la concatenación, se puede completar una columna que identifique categóricamente los datos de muestra como provenientes del primer archivo (muestra de prueba) o del segundo archivo (muestra de control).

[0061] En el bloque 302, el sistema selecciona subconjuntos del archivo concatenado en respuesta a una entrada del usuario. Esto podría ser el dibujo de una ventana o múltiples ventanas usando una interfaz, o la división en subconjuntos de muestras y/o sus datos consiguientes en función de variables categóricas, por ejemplo, el estado de la enfermedad. Los valores pueden recibirse desde una interfaz de usuario y ser procesados por el sistema para identificar los subconjuntos apropiados. Por ejemplo, cuando un usuario dibuja una ventana en un gráfico presentado a través de una interfaz de usuario, los eventos incluidos dentro de la ventana pueden asociarse con un subconjunto del archivo concatenado.

[0062] En el bloque 306, el sistema realiza operaciones de reducción de datos en dos poblaciones diferentes de un archivo o en un archivo de resumen construido mediante la unión de archivos de datos para su comparación. En ambos casos, la reducción de datos se realiza en un conjunto/matriz de datos en lugar de ejecutarse por separado. Un ejemplo de operación de reducción de datos es la reducción de datos con t-SNE. Ejemplos adicionales de operaciones de reducción de datos que se pueden realizar incluyen el análisis de componentes principales (ACP), el análisis discriminante lineal (ADL) y la alineación del espacio tangente local (LTSA por sus siglas en inglés, *local tangent space alignment*). La operación de reducción de datos produce nuevos parámetros para el primer y segundo conjuntos de datos (por ejemplo, valores tSNE para las células en la tabla 200). Los datos multidimensionales están entonces listos para el análisis comparativo a partir del bloque 308, como se examina a continuación. Si el archivo de datos ya tiene al menos parámetros que son el resultado de la reducción de datos, entonces no es necesario realizar el bloque 306.

[0063] En una segunda opción, un usuario identifica ventanas o poblaciones dentro de una muestra en el bloque 304. Esto permite al usuario analizar comparativamente diferentes poblaciones dentro de una única muestra (a diferencia del análisis entre muestras diferentes, como ocurre con los bloques 302-304). A continuación, se puede realizar el bloque 306 después de que se hayan identificado las ventanas/poblaciones, lo que produce datos multidimensionales listos para el análisis comparativo a partir del bloque 308.

[0064] En el bloque 308, el procesador selecciona n subconjuntos de los datos de n dimensiones para su comparación en respuesta a una entrada del usuario. Como ejemplo, n puede ser 2, definiendo así el Subconjunto A y el Subconjunto B. Estos subconjuntos pueden corresponder al primer y segundo conjunto de datos que se van a evaluar comparativamente mediante el acotamiento de zonas de interés por diferencia de frecuencia. Por ejemplo, se pueden realizar las selecciones de subconjuntos basándose en variables categóricas como el parámetro que identifica si los datos de muestra son para poblaciones de prueba o poblaciones de control (por ejemplo, poblaciones con cáncer frente a poblaciones sanas). Sin embargo, debe entenderse que estos subconjuntos se pueden definir basándose en cualquier parámetro en los datos de n dimensiones (por ejemplo, parámetro de tejido canceroso frente a parámetro de tejido sano producido por el bloque 300).

[0065] A continuación, en el bloque 310, el procesador selecciona n parámetros de los Subconjuntos A y B para definir la base para comparar los Subconjuntos A y B. Como ejemplo, n puede ser 2. Los parámetros seleccionados pueden ser parámetros presentes en los datos de n dimensiones, parámetros derivados de los datos de n dimensiones y/o parámetros creados por otros enfoques de reducción de datos. Las distribuciones pueden provenir de diferentes muestras, pero también de subconjuntos de la misma muestra que comparten n parámetros para su comparación.

[0066] A continuación, en el bloque 311, el procesador genera una estimación de frecuencia bivariada que se realiza calculando un histograma bidimensional para cada muestra de comparación. El histograma se normaliza

mediante el recuento de eventos y, por lo general, aunque no necesariamente, se suaviza utilizando un suavizado de núcleo de ancho variable que es el mismo que se utiliza para generar gráficos de contorno suavizado o pseudocolor.

[0067] En el bloque 312, el procesador calcula dos histogramas de diferencia para cada elemento (por ejemplo, intervalo o *bin*) en el histograma. Los valores positivos indican que la región tiene más eventos en el primer comparador; los valores negativos indican que la región tiene más eventos en el segundo comparador.

[0068] A continuación, en el bloque 313, el histograma de diferencia se binormaliza, en donde los valores mayores de 0, correspondientes a la región que tiene más eventos en el primer comparador, se reescalan de 0 a 100 (de modo que la mayor diferencia en el histograma ahora es 100). Los valores inferiores a 0, correspondientes a la región que tiene más eventos en el segundo comparador, se reescalan de manera similar de 0 a -100. Deberá entenderse que en un análisis por lotes donde se generan y comparan múltiples histogramas de diferencia, el usuario puede seleccionar un factor de reescalado global positivo y negativo para aplicar a todos los histogramas con el fin de obtener una mejor comparabilidad.

[0069] A continuación, en el bloque 314, se dibuja el histograma resultante usando una representación de mapa de calor (asignando colores al grado de diferencia), pero deberá entenderse que esto puede incluir la representación usando otros tipos de visualización a través de un dispositivo de visualización. En el bloque 314, el procesador genera una visualización que permite visualizar las diferencias entre los Subconjuntos A y B según una distribución bivariada. Se trata de una nueva y potente visualización que proporciona a los usuarios nuevos conocimientos sobre conjuntos de datos multiparamétricos que no estaban disponibles con los sistemas convencionales en la técnica. Esta visualización proporciona una superposición de las poblaciones elegidas para los Subconjuntos A y B. Esta superposición se puede codificar por colores para indicar visualmente las áreas pobladas con mayor frecuencia del Subconjunto A en relación con el Subconjunto B (por ejemplo, Color 1) y las áreas pobladas con mayor frecuencia del Subconjunto B en relación con el subconjunto A (por ejemplo, Color 2).

[0070] La Figura 4A muestra un ejemplo de dicha visualización. La Figura 4A muestra un gráfico que superpone dos muestras (Subconjunto A: sangre de DS; Subconjunto B: Sangre del paciente) en el espacio de parámetros (t-SNE P 1/2 frente a t-SNE P 2/2). Las regiones (por ejemplo, 500 en la Figura 4A) en el gráfico donde el acotamiento de zonas de interés por diferencia de frecuencia revela una mayor frecuencia de eventos en el Subconjunto A en relación con el Subconjunto B de acuerdo con un umbral definido se muestran en un primer color/sombreado (por ejemplo, azul), y las regiones (por ejemplo, 502 en la Figura 4A) en el gráfico donde el acotamiento de zonas de interés por diferencia de frecuencia revela una mayor frecuencia de eventos en el Subconjunto B en relación con el Subconjunto A de acuerdo con un umbral definido se muestran en un segundo color/sombreado (por ejemplo, rojo). La codificación por colores puede modular la intensidad de la coloración/sombreado en función de la magnitud de las diferencias de frecuencia, como lo indica la leyenda 504 de la visualización.

[0071] Los umbrales definidos para el acotamiento de zonas de interés por diferencia de frecuencia pueden ser umbrales fijos o pueden ser umbrales ajustables. Por ejemplo, la visualización puede ser una visualización interactiva donde un usuario puede ajustar el umbral o los umbrales definidos a través de las entradas 510 y 512. En el ejemplo de la Figura 4A, el usuario puede definir un umbral para el límite de ventana alto (más frecuente en el Subconjunto A que en el Subconjunto B) a través de la entrada 510. Específicamente, para el acotamiento de zonas de interés (selección de subconjunto), el usuario introduce un rango de valores de diferencia para incluir en la región. Por ejemplo, de 0 a 100 (máximo) seleccionará todas las regiones donde los eventos son más frecuentes en el primer comparador. Se pueden utilizar valores más estrictos para seleccionar regiones con mayor diferencia. En el ejemplo de la Figura 4A, el usuario también puede definir un umbral para el límite de ventana bajo (más frecuente en el Subconjunto B que en el Subconjunto A) a través de la entrada 512. Sin embargo, también deberá entenderse que se puede usar un umbral único, en cuyo caso el acotamiento de zonas de interés es una elección binaria entre “Más frecuencia en el Subconjunto A” o “Más frecuencia en el Subconjunto B”, aunque los inventores creen que múltiples umbrales, como se muestra en la Figura 4A, pueden proporcionar conocimientos más profundos sobre las propiedades biológicas de los datos.

[0072] Tomando como base esta visualización, el usuario puede elegir si desea crear cualesquiera ventanas en función de las diferencias de frecuencia presentadas en el bloque 315 de la Figura 3. El área de entrada del usuario 514 permite al usuario identificar qué ventana, de entre una pluralidad de ventanas, se puede crear a partir de los datos tras la selección del botón “Crear ventanas” 516. Las opciones en el área 514 incluyen (1) una opción de “Crear ventana superior”, que acota las regiones donde los eventos son más frecuentes en el primer comparador (0 a 100), configurada usando el campo 510, pudiéndose usar valores más estrictos para seleccionar regiones con mayor diferencia, (2) una opción “Crear ventana inferior”, que acota las regiones donde los eventos son más frecuentes en el segundo comparador (0 a -100), configurada usando 512, pudiéndose usar valores más estrictos para seleccionar regiones con mayor diferencia, y (3) una opción “Crear ventana de rango medio”, que acota las regiones que quedan fuera de las ventanas superior e inferior. Estas diferentes regiones acotadas como zonas de interés pueden proporcionar al usuario diferentes conocimientos sobre los datos porque las regiones que son diferentes de alguna manera (por ejemplo, mayor o menor frecuencia de eventos) o iguales de alguna manera pueden ser biológicamente interesantes para el usuario.

[0073] La Figura 4B muestra un ejemplo de cómo la visualización se puede ajustar interactivamente en función de entradas del usuario. En la Figura 4B, el usuario ha seleccionado la opción de controlar el gráfico para mostrar solo las regiones acotadas definidas, que en este caso es una ventana superior definida donde los eventos son más frecuentes en el primer comparador y es igual a 20 (de 100) y los eventos son más frecuentes en el segundo comparador y se configura en -20 (de -100). Esto produce regiones codificadas por colores/sombreadas (por ejemplo, 500/502), como se muestra en la Figura 4B.

[0074] Ejemplos adicionales de controles interactivos sobre la visualización pueden incluir un control de sensibilidad y un control de especificidad. Las tres estadísticas (especificidad, sensibilidad y valor p) solo se calculan si cada comparador está compuesto por un grupo de sujetos, por ejemplo, el primer comparador está formado por subconjuntos de n sujetos. A continuación, se pueden calcular las estadísticas para determinar qué fracción de eventos en cada sujeto están incluidos en la región seleccionada. Estos se utilizan para cálculos de sensibilidad y especificidad. Se calcula el valor P, aunque no se muestra, que es la prueba T de Student sobre la fracción de eventos en la ventana para los sujetos del grupo 1 frente a los sujetos del grupo 2. El control de sensibilidad puede controlar qué fracción de los eventos en la población comparada aparecería en la ventana cuando se crea, como se muestra en la Figura 4C como control deslizante 520. El control de especificidad puede controlar qué fracción de los eventos en la ventana provienen de la población comparada, por ejemplo, la “pureza” de la ventana, mostrada en la Figura 4C como control deslizante 521.

[0075] Tras la selección de opciones de acotamiento de zonas de interés o ventanas en el área 514 y el botón 516, el sistema genera un conjunto de datos correspondiente a las ventanas definidas (bloque 320). Estas ventanas definidas luego se pueden agrupar o dividir en subconjuntos para explorar todas las diferencias entre n muestras. Además, se pueden explorar las ventanas en busca de otros parámetros para la expresión de proteínas de genes a fin de identificar los subconjuntos de células que componen estas poblaciones. Se pueden añadir los subconjuntos de células/eventos definidos por la ventana creada a la población de control.

[0076] Por lo tanto, deberá entenderse que las técnicas de acotamiento de zonas de interés por diferencia de frecuencia descritas en el presente proporcionan a los usuarios una poderosa herramienta que explora distribuciones multivariadas complejas y cuantifica las diferencias entre muestras basándose en múltiples mediciones. Una herramienta de este tipo permite a los usuarios generar información sobre grandes conjuntos multiparamétricos que no están disponibles en los sistemas convencionales en la técnica. Por ejemplo, el acotamiento de zonas de interés por diferencia de frecuencia proporciona una herramienta imparcial para identificar rápidamente regiones en el espacio multivariado en las que la frecuencia de eventos es estadísticamente significativamente diferente entre muestras. Se pueden utilizar estas regiones identificadas de varias maneras útiles, que incluyen, entre otras: (1) la identificación de células que responden a un estímulo; y (2) la identificación de diferencias asociadas a enfermedades en el fenotipo o la representación. Asimismo, se puede aplicar el acotamiento de zonas de interés por diferencia de frecuencia a otras muestras para cuantificar el número de “respondedores”.

[0077] A través de estas y otras características, los ejemplos de realizaciones de la invención proporcionan avances técnicos significativos en el campo de la bioinformática aplicada.

[0078] Tal y como se usan en el presente, los términos establecidos con particularidad más adelante tienen las siguientes definiciones. Si no se define de otro modo en esta sección, todos los términos utilizados en el presente tienen el significado comúnmente entendido por un experto en la materia a la que pertenece esta invención.

[0079] Tal y como se usan en el presente, “sistema”, “instrumento”, “aparato” y “dispositivo” generalmente abarcan los componentes de *hardware* (por ejemplo, mecánicos y electrónicos) y, en algunas implementaciones, los componentes de *software* asociados (por ejemplo, programas informáticos especializados para control de gráficos).

[0080] Tal y como se usa en el presente, un “evento” generalmente hace referencia al paquete de datos medidos a partir de una sola partícula, por ejemplo, células o partículas sintéticas. Normalmente, los datos medidos a partir de una sola partícula incluyen una serie de parámetros, incluidos uno o más parámetros de dispersión de la luz, y al menos un parámetro o característica derivado de la fluorescencia y detectado a partir de la partícula, tal como la intensidad de fluorescencia. Por consiguiente, cada evento se representa como un vector de mediciones y características, en donde cada parámetro o característica medidos corresponde a una dimensión del espacio de datos. En algunas realizaciones, los datos medidos de una sola partícula pueden incluir datos de imagen, eléctricos, temporales o acústicos. En algunas aplicaciones biológicas, los datos de eventos pueden corresponder a datos biológicos cuantitativos que indican la expresión de una proteína o gen específicos.

[0081] Tal y como se usan en el presente, una “población” o “subpoblación” de partículas, como por ejemplo células u otras partículas, generalmente hacen referencia a un grupo de partículas que poseen propiedades (por ejemplo, propiedades ópticas, de impedancia o temporales) con respecto a uno o más parámetros medidos, de tal manera que los datos de parámetros medidos forman un clúster en el espacio de datos. Por consiguiente, las poblaciones se reconocen como clústeres en los datos. A la inversa, cada clúster de datos generalmente se interpreta como que

corresponde a una población de un tipo particular de célula o partícula, aunque también se observan normalmente clústeres que corresponden a ruido o al fondo. Un clúster puede definirse en un subconjunto de las dimensiones, es decir, con respecto a un subconjunto de los parámetros medidos, que corresponde a poblaciones que difieren en solo un subconjunto de los parámetros o características medidos extraídos de las mediciones de la célula o partícula.

[0082] Como se usa en el presente, una "ventana" (*gate*) generalmente hace referencia a un límite clasificador que identifica un subconjunto de datos de interés. En citometría, una ventana puede delimitar un grupo de eventos de particular interés. Tal y como se usa en el presente, el "acotamiento de zonas de interés" (*gating* en inglés) generalmente hace referencia al proceso de clasificar los datos usando una ventana definida para un conjunto de datos determinado, donde la ventana puede ser una o más regiones de interés combinadas, en algunos casos, utilizando lógica booleana.

[0083] Tal y como se usa en el presente, un "evento" generalmente hace referencia al paquete reunido de datos medidos de una sola partícula (por ejemplo, células o partículas sintéticas). Normalmente, los datos medidos a partir de una sola partícula incluyen una serie de parámetros o características, incluidos uno o más parámetros o características de dispersión de la luz, y al menos otro parámetro o característica derivados de fluorescencia medida. Por consiguiente, cada evento se representa como un vector de mediciones de parámetros o características, en donde cada parámetro o característica medidos corresponde a una dimensión del espacio de datos.

[0084] Tal y como se usan en el presente, los términos "determinar", "determinando", "determinación" o "que determinan" abarcan una amplia variedad de acciones. Por ejemplo, el término "determinar" puede incluir calcular, computar, procesar, derivar, investigar, buscar (por ejemplo, buscar en una tabla, en una base de datos o en otra estructura de datos), averiguar y similares. Además, el término "determinar" puede incluir recibir (por ejemplo, recibir información), acceder (por ejemplo, acceder a datos en una memoria) y similares. Asimismo, "determinar" puede incluir resolver, seleccionar, elegir, establecer y acciones similares.

[0085] Tal y como se usan en el presente, los términos "proporcionar", "proporcionando" o "que proporcionan" abarcan una amplia variedad de acciones. Por ejemplo, el término "proporcionar" puede incluir almacenar un valor en una ubicación para su posterior recuperación, transmitir un valor directamente al destinatario, transmitir o almacenar una referencia a un valor y similares. "Proporcionar" también puede incluir codificar, decodificar, cifrar, descifrar, validar, verificar y acciones similares.

[0086] Tal y como se usa en el presente, el término "selectivamente" o "selectivo/a" puede abarcar una amplia variedad de acciones. Por ejemplo, un proceso "selectivo" puede incluir la determinación de una opción entre múltiples opciones. Un proceso "selectivo" puede incluir uno o más de: entradas determinadas dinámicamente, entradas preconfiguradas o entradas iniciadas por el usuario para tomar la determinación. En algunas implementaciones se puede incluir un interruptor de "n" entradas para proporcionar una funcionalidad selectiva en donde "n" es el número de entradas utilizadas para realizar la selección.

[0087] Tal y como se usa en el presente, el término "mensaje" abarca una amplia variedad de formatos para comunicar (por ejemplo, transmitir o recibir) información. Un mensaje puede incluir una agregación de información legible por máquina, tal como un documento XML, un mensaje de campo fijo, un mensaje separado por comas o similares. Un mensaje puede, en algunas implementaciones, incluir una señal utilizada para transmitir una o más representaciones de la información. Aunque se presente en singular, se entenderá que un mensaje puede componerse, transmitirse, almacenarse, recibirse, etc. en múltiples partes.

[0088] Tal y como se usa en el presente, una expresión que hace referencia a "al menos uno/a de" una lista de elementos se refiere a cualquier combinación de esos elementos, incluidos los miembros individuales. Por ejemplo, "al menos uno de: a, b o c" tiene como objetivo abarcar: a, b, c, a-b, a-c, b-c y a-b-c.

[0089] Los expertos en la materia entenderán que la información, los mensajes y las señales pueden representarse utilizando cualquiera de una variedad de tecnologías y técnicas diferentes. Por ejemplo, los datos, instrucciones, comandos, información, señales, bits, símbolos y chips a los que se puede hacer referencia en toda la descripción anterior pueden estar representados por voltajes, corrientes, ondas electromagnéticas, partículas o campos magnéticos, partículas o campos ópticos o cualquier combinación de los mismos.

[0090] Los expertos en la materia apreciarán además que los diversos bloques lógicos, módulos, circuitos y pasos de algoritmo ilustrativos descritos en relación con las realizaciones dadas a conocer en el presente pueden implementarse como hardware electrónico, software informático o combinaciones de ambos. Para ilustrar claramente esta intercambiabilidad de hardware y software, varios componentes, bloques, módulos, circuitos y pasos ilustrativos han sido descritos anteriormente generalmente en términos de su funcionalidad. Que dicha funcionalidad se implemente como hardware o software depende de la aplicación específica y de las restricciones de diseño impuestas en el sistema general. Los expertos en la materia pueden implementar la funcionalidad descrita de diferentes maneras para cada aplicación específica, pero estas decisiones de implementación no deben interpretarse como causas de una desviación con respecto al ámbito de la presente invención.

[0091] Las técnicas descritas en el presente pueden implementarse en *hardware*, *software*, *firmware* o cualquier combinación de los mismos. Dichas técnicas pueden implementarse en cualquiera de una variedad de dispositivos, como por ejemplo ordenadores de procesamiento de eventos programados específicamente, dispositivos de comunicación inalámbrica o dispositivos de circuitos integrados. Todas las características descritas como módulos o componentes pueden implementarse conjuntamente en un dispositivo lógico integrado o independientemente como dispositivos lógicos discretos pero interoperables. Si se implementan en software, las técnicas pueden realizarse al menos en parte mediante un medio de almacenamiento de datos legible por ordenador que comprende un código de programa que incluye instrucciones que, cuando se ejecutan, llevan a cabo uno o más de los métodos descritos anteriormente. El medio de almacenamiento de datos legible por ordenador puede formar parte de un producto de programa informático, el cual puede incluir materiales de embalaje. El medio legible por ordenador puede comprender memoria o medios de almacenamiento de datos, como por ejemplo memoria de acceso aleatorio (RAM por sus siglas en inglés, *Random Access Memory*), memoria de acceso aleatorio dinámico síncrono (SDRAM por sus siglas en inglés, *Synchronous Dynamic Random-Access Memory*), memoria de solo lectura (ROM por sus siglas en inglés, *Read-Only Memory*), memoria de acceso aleatorio no volátil (NVRAM por sus siglas en inglés, *Non-Volatile Random Access Memory*), memoria de solo lectura programable y borrable eléctricamente (EEPROM por sus siglas en inglés, *Electrically Erasable Programmable Read-Only Memory*), memoria flash, medios de almacenamiento de datos magnéticos u ópticos, y similares. El medio legible por ordenador puede ser un medio de almacenamiento no transitorio. Las técnicas, adicional o alternativamente, pueden implementarse al menos en parte mediante un medio de comunicación legible por ordenador que transporta o comunica el código de programa en forma de instrucciones o estructuras de datos y que un ordenador puede acceder, leer y/o ejecutar, como por ejemplo señales u ondas propagadas.

[0092] El código de programa puede ejecutarse mediante un procesador de gráficos programado específicamente, que puede incluir uno o más procesadores, como por ejemplo uno o más procesadores de señales digitales (DSP por sus siglas en inglés, *Digital Signal Processors*), microprocesadores configurables, circuitos integrados para aplicaciones específicas (ASIC por sus siglas en inglés, *Application-Specific Integrated Circuits*), matrices de puertas lógicas programables en campo (FPGA por sus siglas en inglés, *Field Programmable Gate Arrays*) u otros circuitos lógicos integrados o discretos equivalentes. Un procesador de gráficos de este tipo puede configurarse especialmente para realizar cualquiera de las técnicas descritas en esta divulgación. Una combinación de dispositivos informáticos, por ejemplo, una combinación de un DSP y un microprocesador, una pluralidad de microprocesadores, uno o más microprocesadores junto con un núcleo DSP, o cualquier otra configuración similar en conectividad de datos al menos parcial puede implementar uno o más de los elementos descritos. En consecuencia, el término "procesador", tal y como se utiliza en el presente, puede referirse a cualquiera de las estructuras anteriores, a cualquier combinación de las estructuras anteriores o a cualquier otra estructura o aparato adecuados para la implementación de las técnicas descritas en el presente. Asimismo, en algunos aspectos, la funcionalidad descrita en el presente puede proporcionarse dentro de módulos de software dedicados o módulos de hardware configurados para codificar y decodificar, o incorporarse en una tarjeta de control gráfico especializada.

[0093] Los métodos descritos en el presente comprenden uno o más pasos o acciones para realizar el método descrito. Los pasos y/o acciones del método pueden intercambiarse entre sí sin abandonar el ámbito de las reivindicaciones. En otras palabras, a menos que se especifique un orden concreto de pasos o acciones, el orden y/o el uso de pasos y/o acciones específicos pueden modificarse sin abandonar el ámbito de las reivindicaciones.

[0094] Se han descrito diversas realizaciones de la invención. Estas y otras realizaciones se encuentran dentro del ámbito de las reivindicaciones que se muestran a continuación.

REIVINDICACIONES

1. Un método implementado por ordenador para visualizar diferencias entre conjuntos de datos de n dimensiones, comprendiendo este método implementado por ordenador:
bajo el control de uno o más dispositivos de procesamiento,
la concatenación de un primer conjunto de datos de n dimensiones y un segundo conjunto de datos de n dimensiones para obtener un conjunto de datos concatenados, en donde los datos de n dimensiones comprenden una pluralidad de eventos en una pluralidad de dimensiones;
la ejecución de una reducción de dimensionalidad en el conjunto de datos concatenados para obtener una asignación de eventos en el primer y segundo conjunto de datos:
la ejecución de un acotamiento de zonas de interés (*gating*) por diferencia de frecuencia en la asignación de eventos mediante:
la selección de un primer y de un segundo subconjunto de datos de n dimensiones para compararlos en respuesta a la entrada del usuario;
la generación de un histograma multidimensional con una pluralidad de elementos por dimensión de acuerdo con una estimación de frecuencia bivariada de cada uno del primer subconjunto y el segundo subconjunto dentro del primer conjunto de datos y el segundo conjunto de datos, comprendiendo el histograma multidimensional una pluralidad de elementos;
la normalización del histograma multidimensional por un recuento de eventos;
la generación de histogramas de diferencia para cada elemento del histograma multidimensional; y
la binormalización de los histogramas de diferencia, en donde los valores correspondientes a una región que tiene más eventos en el primer subconjunto se reescalan mediante un factor de reescalado positivo global, y los valores correspondientes a una región que tiene más eventos en el segundo subconjunto se reescalan mediante un factor de reescalado negativo global;
la generación de una visualización a partir del acotamiento de zonas de interés por diferencia de frecuencia para su visualización a través de un dispositivo de visualización, mostrando la visualización regiones en un espacio de variables múltiples donde una frecuencia de eventos del primer subconjunto es diferente de una frecuencia de eventos del segundo subconjunto de acuerdo con un umbral definido.

2. El método implementado por ordenador de la Reivindicación 1, en donde el primer y segundo conjuntos de datos comprenden datos de muestras de células multiparamétricas.

3. El método implementado por ordenador de la Reivindicación 1, que además comprende:
el ajuste del umbral definido en respuesta a una entrada del usuario; y
el ajuste de la visualización en función del umbral definido ajustado.

4. El método implementado por ordenador de la Reivindicación 1, en donde el umbral definido comprende una pluralidad de umbrales definidos.

5. El método implementado por ordenador de la Reivindicación 1, en donde la generación de la visualización comprende la codificación por colores de las regiones basándose al menos en parte en el acotamiento de zonas de interés por diferencia de frecuencia.

6. El método implementado por ordenador de la Reivindicación 1, en donde el umbral definido incluye un umbral superior que identifica una o más regiones clasificadas por tener una mayor frecuencia de eventos del primer subconjunto que del segundo subconjunto.

7. El método implementado por ordenador de la Reivindicación 1, en donde el umbral definido incluye un umbral inferior que identifica una o más regiones que tienen una mayor frecuencia de eventos del segundo subconjunto que del primer subconjunto.

8. El método implementado por ordenador de la Reivindicación 1, en donde el umbral definido incluye límites de rango medio que identifican una o más regiones que tienen una frecuencia similar de eventos entre el primer subconjunto y el segundo subconjunto.

9. El método implementado por ordenador de la Reivindicación 1, en donde la generación de la visualización comprende: la representación de un mapa de calor de los histogramas de diferencia binormalizados.

10. El método implementado por ordenador de la Reivindicación 1, que además comprende:
la generación de un tercer conjunto de datos basado en al menos una ventana definida por un usuario a través de la visualización.

11. El método implementado por ordenador de la Reivindicación 1, en donde el primer conjunto de datos comprende una muestra de control.

12. El método implementado por ordenador de la Reivindicación 11, en donde la muestra de control corresponde a datos celulares procedentes de materia sana.

13. El método implementado por ordenador de la Reivindicación 11, en donde la muestra de control corresponde a datos celulares procedentes de materia cancerosa.

14. Un sistema que comprende:
uno o más dispositivos de procesamiento; y
un medio de almacenamiento legible por ordenador que comprende instrucciones que, cuando son ejecutadas por uno o más dispositivos de procesamiento, hacen que el sistema:
reciba un umbral para el acotamiento de zonas de interés por diferencia de frecuencia;
reciba un primer conjunto de datos de n dimensiones que incluye una primera pluralidad de eventos en una pluralidad de dimensiones;
reciba un segundo conjunto de datos de n dimensiones que incluye una segunda pluralidad de eventos en al menos la pluralidad de dimensiones;
concatene el primer conjunto de datos y el segundo conjunto de datos para obtener un conjunto de datos concatenados;
ejecute una reducción de dimensionalidad en el conjunto de datos concatenados para obtener una asignación de eventos en el primer y segundo conjuntos de datos;
identifique una ventana de diferencia de frecuencia que define una población de eventos basada, al menos en parte, en el acotamiento de zonas de interés por diferencia de frecuencia realizado en la asignación de eventos, identificando esta ventana una región en el espacio multivariado donde una frecuencia de eventos de un primer subconjunto de datos de n dimensiones es diferente de una frecuencia de eventos de un segundo subconjunto de datos de n dimensiones, de conformidad con el umbral, en donde el acotamiento de zonas de interés por diferencia de frecuencia comprende:
la selección de dichos primer y segundo subconjuntos de datos de n dimensiones para compararlos en respuesta a una entrada del usuario;
la generación de un histograma multidimensional con una pluralidad de elementos por dimensión de acuerdo con una estimación de frecuencia bivariada de cada uno del primer subconjunto y el segundo subconjunto dentro del primer conjunto de datos y el segundo conjunto de datos, comprendiendo el histograma multidimensional una pluralidad de elementos;
la normalización del histograma multidimensional por un recuento de eventos;
la generación de histogramas de diferencia para cada elemento del histograma multidimensional; y
la binormalización de los histogramas de diferencia en donde los valores correspondientes a una región que tiene más eventos en el primer subconjunto se reescalan mediante un factor de reescalado positivo global, y los valores correspondientes a una región que tiene más eventos en el segundo subconjunto se reescalan mediante un factor de reescalado negativo global; y
hacen que se muestre una visualización que incluye una representación de eventos del primer conjunto de datos y del segundo conjunto de datos incluidos en la población definida por la ventana de diferencia de frecuencia, mostrando esta visualización regiones en el espacio multivariado donde la frecuencia de eventos del primer subconjunto es diferente de la frecuencia de eventos del segundo subconjunto de conformidad con el umbral.

15. El sistema de la Reivindicación 14, en donde el umbral incluye al menos uno de los siguientes elementos:
un umbral superior que identifica una o más regiones clasificadas como que tienen una mayor frecuencia de eventos del primer subconjunto que del segundo subconjunto;
un umbral inferior que identifica una o más regiones que tienen una mayor frecuencia de eventos del segundo subconjunto que del primer subconjunto; o
límites de rango medio que identifican una o más regiones que tienen una frecuencia similar de eventos entre el primer subconjunto y el segundo subconjunto.

16. El sistema de la Reivindicación 14, en donde los elementos del histograma multidimensional son intervalos o *bins*.

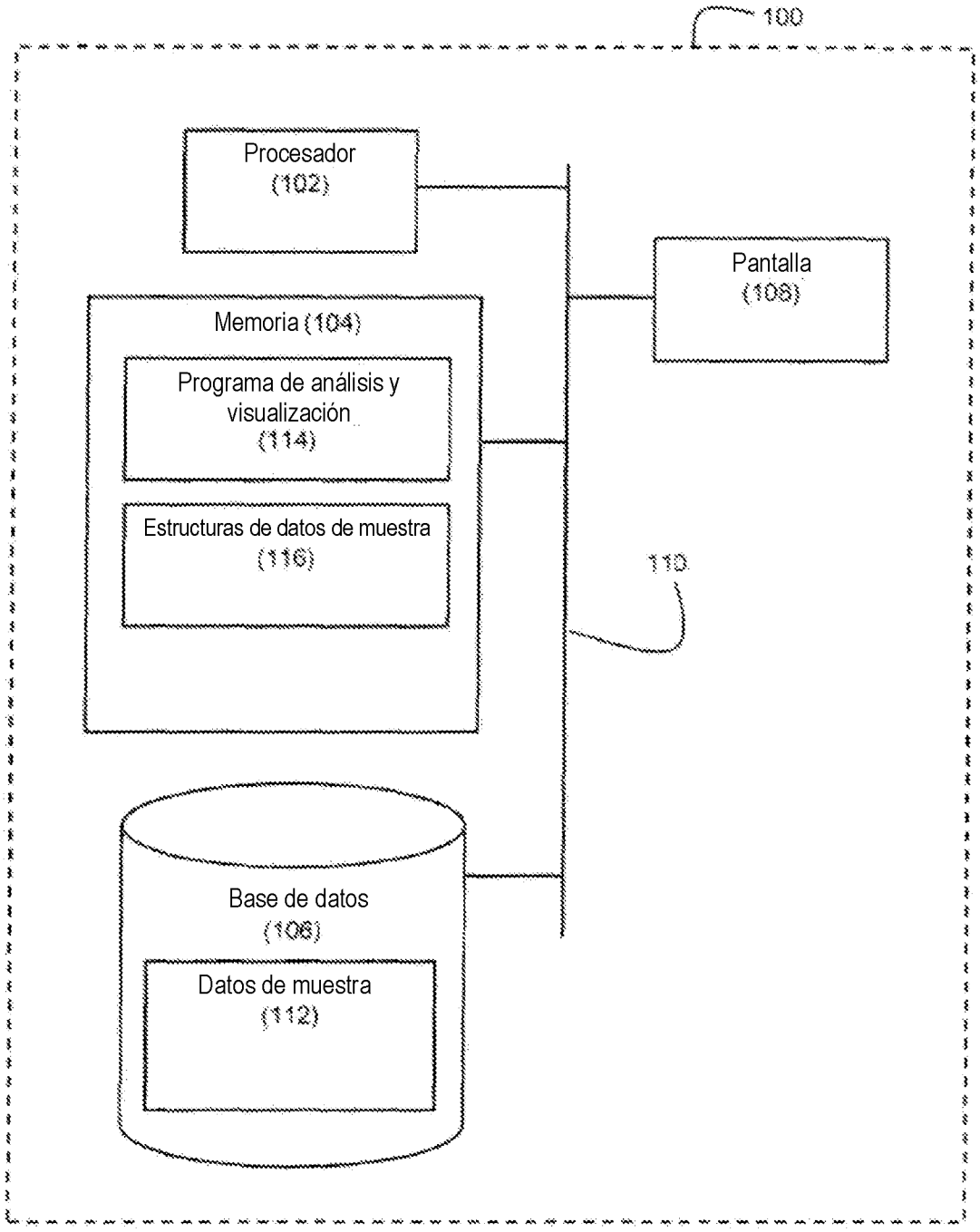
100
Procesador
(102)
Pantalla
(108)
Memoria (104)
Programa de análisis y visualización
(114)
Estructuras de datos de muestra
(116)
110
Base de datos
(106)
Datos de muestra
(112)

Figura 1

Figura 2A

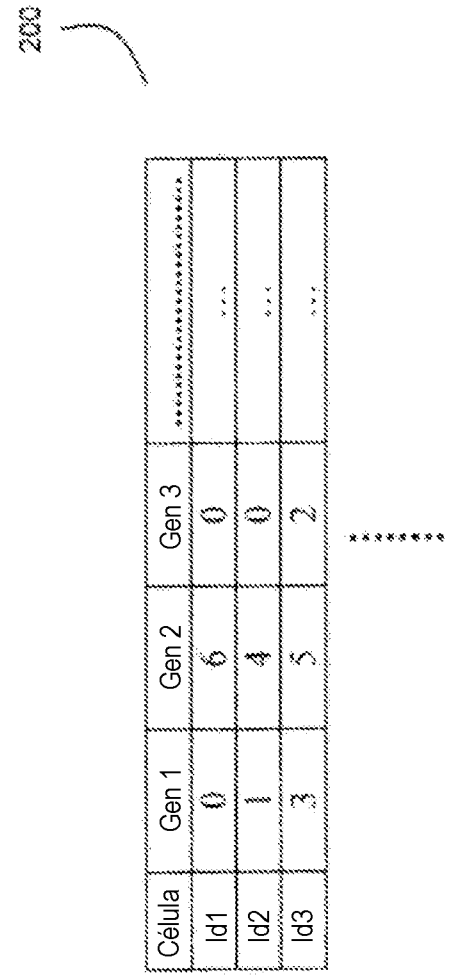

Figura 2B

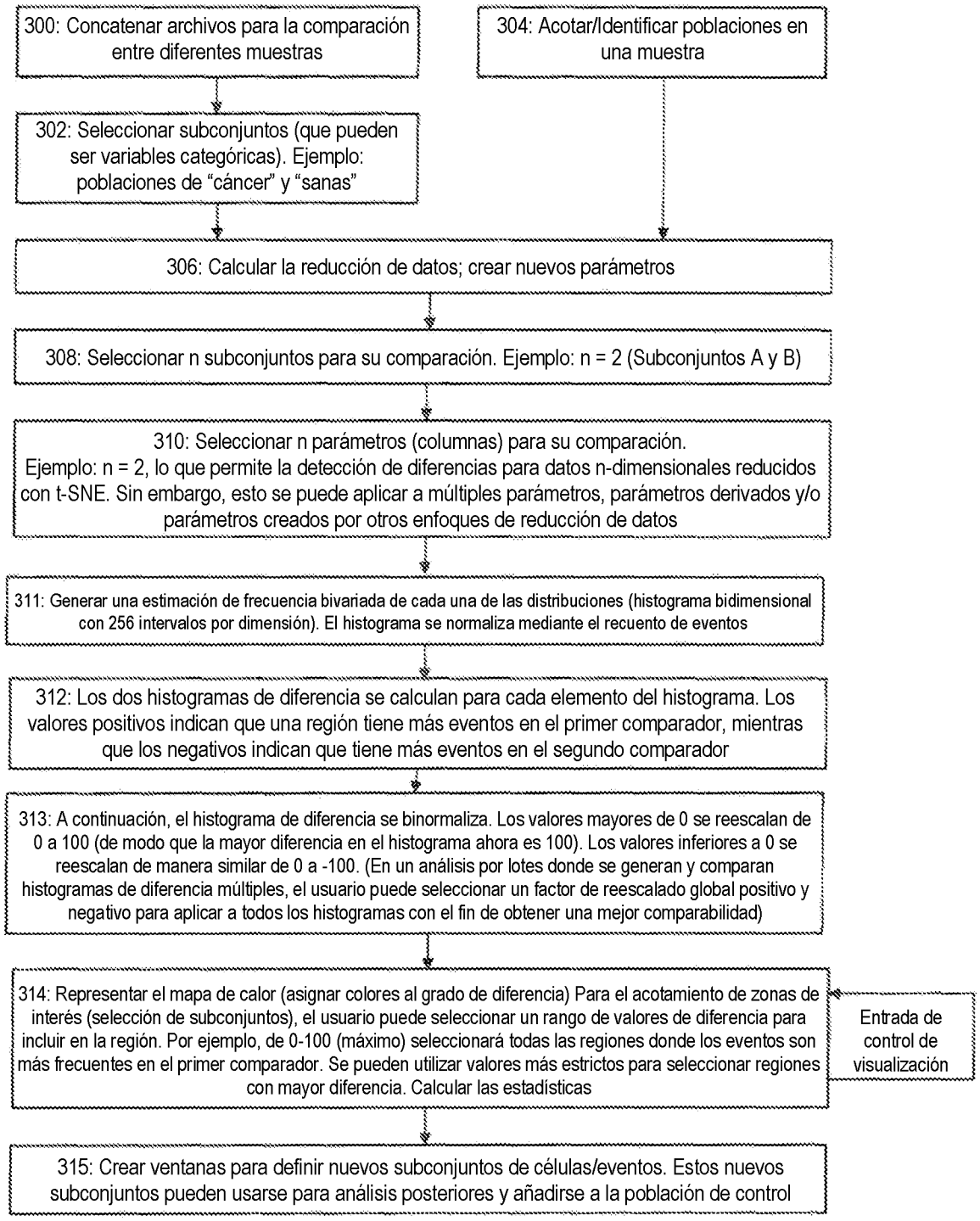
300: Concatenar archivos para la comparación entre diferentes muestras
304: Acotar/Identificar poblaciones en una muestra
302: Seleccionar subconjuntos (que pueden ser variables categóricas). Ejemplo: poblaciones de "cáncer" y "sanas"
306: Calcular la reducción de datos; crear nuevos parámetros
308: Seleccionar n subconjuntos para su comparación. Ejemplo: n = 2 (Subconjuntos A y B)
310: Seleccionar n parámetros (columnas) para su comparación. Ejemplo: n = 2, lo que permite la detección de diferencias para datos n-dimensionales reducidos con t-SNE. Sin embargo, esto se puede aplicar a múltiples parámetros, parámetros derivados y/o parámetros creados por otros enfoques de reducción de datos
311: Generar una estimación de frecuencia bivariada de cada una de las distribuciones (histograma bidimensional con 256 intervalos por dimensión). El histograma se normaliza mediante el recuento de eventos
312: Los dos histogramas de diferencia se calculan para cada elemento del histograma. Los valores positivos indican que una región tiene más eventos en el primer comparador, mientras que los negativos indican que tiene más eventos en el segundo comparador
313: A continuación, el histograma de diferencia se binormaliza. Los valores mayores de 0 se reescalan de 0 a 100 (de modo que la mayor diferencia en el histograma ahora es 100). Los valores inferiores a 0 se reescalan de manera similar de 0 a -100. (En un análisis por lotes donde se generan y comparan histogramas de diferencia múltiples, el usuario puede seleccionar un factor de reescalado global positivo y negativo para aplicar a todos los histogramas con el fin de obtener una mejor comparabilidad)
314: Representar el mapa de calor (asignar colores al grado de diferencia) Para el acotamiento de zonas de interés (selección de subconjuntos), el usuario puede seleccionar un rango de valores de diferencia para incluir en la región. Por ejemplo, de 0-100 (máximo) seleccionará todas las regiones donde los eventos son más frecuentes en el primer comparador. Se pueden utilizar valores más estrictos para seleccionar regiones con mayor diferencia. Calcular las estadísticas
Entrada de control de visualización
315: Crear ventanas para definir nuevos subconjuntos de células/eventos. Estos nuevos subconjuntos pueden usarse para análisis posteriores y añadirse a la población de control

Figura 3

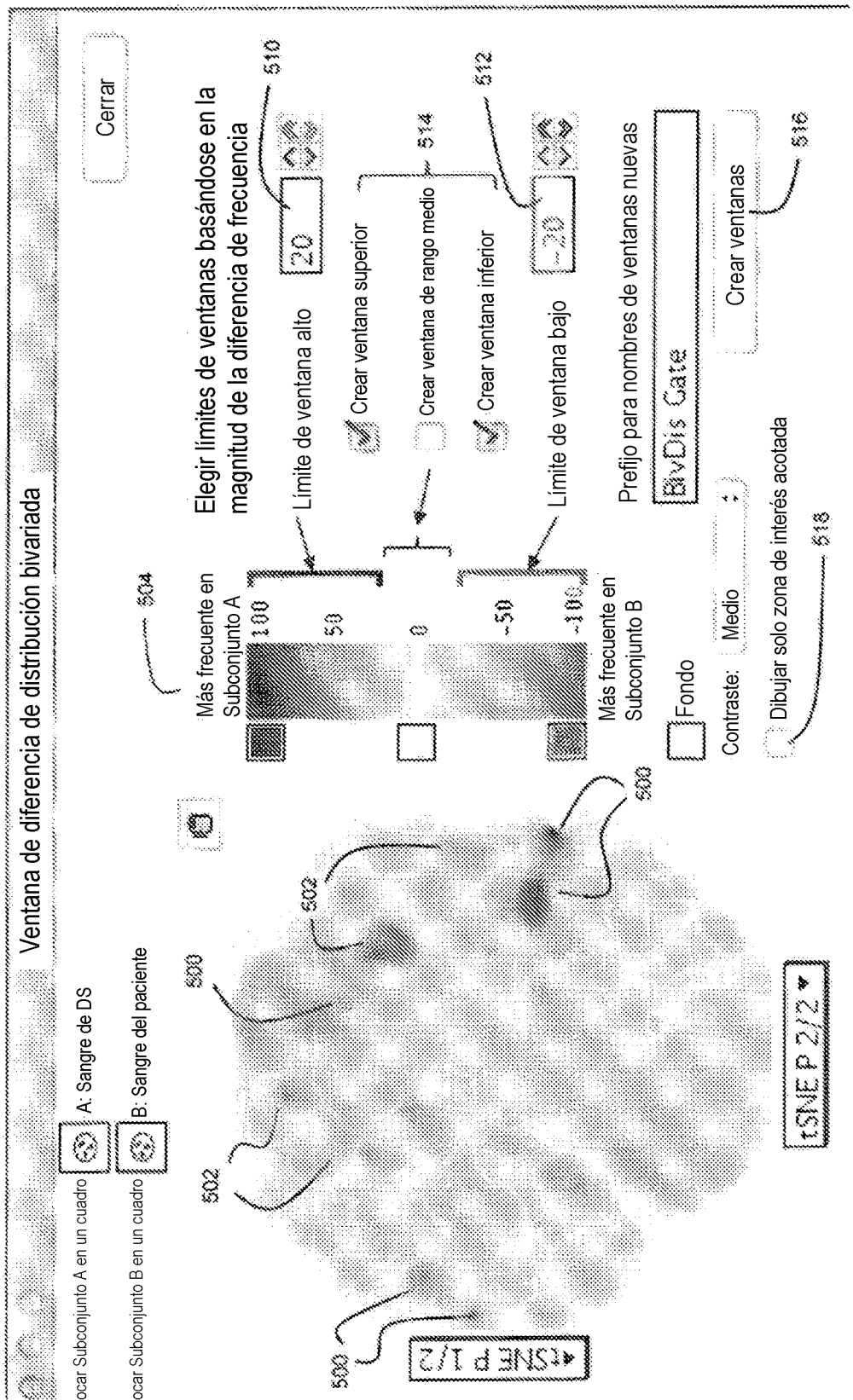
Ventana de diferencia de distribución bivariada
Cerrar
Colocar Subconjunto A en un cuadro
A: Sangre de DS
Colocar Subconjunto B en un cuadro
B: Sangre del paciente
Elegir límites de ventanas basándose en la magnitud de la diferencia de frecuencia
Límite de ventana alto
Crear ventana superior
Crear ventana de rango medio
Crear ventana inferior
Límite de ventana bajo
Más frecuente en Subconjunto A
Más frecuente en Subconjunto B
100
50
0
-50
-100
20
-20
Prefijo para nombres de ventanas nuevas
BivDis Gate
Fondo
Contraste:
Medio
Crear ventanas
Dibujar solo zona de interés acotada
tSNE P 1/2
tSNE P 2/2
500
502
504
506
508
510
512
514
516
518

Figura 4A

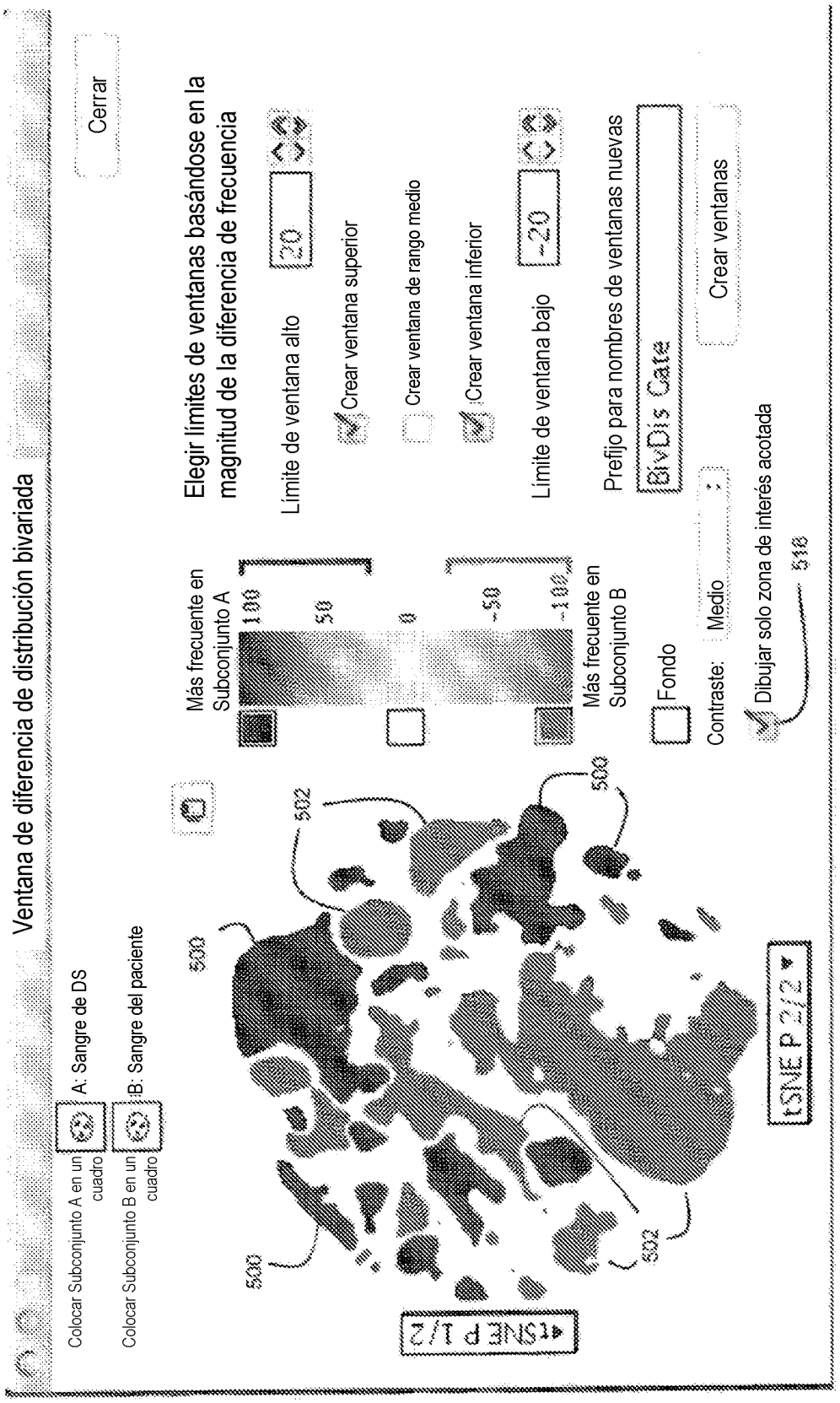
Ventana de diferencia de distribución bivariada
Cerrar
Colocar Subconjunto A en un cuadro
A: Sangre de DS
Colocar Subconjunto B en un cuadro
B: Sangre del paciente
Más frecuente en Subconjunto A
100
50
0
-50
-100
Más frecuente en Subconjunto B
Fondo
Contraste:
Medio
Dibujar solo zona de interés acotada
Elegir límites de ventanas basándose en la magnitud de la diferencia de frecuencia
Límite de ventana alto
20
Crear ventana superior
Crear ventana de rango medio
Crear ventana inferior
Límite de ventana bajo
-20
Prefijo para nombres de ventanas nuevas
BivDis Cate
Crear ventanas
tSNE P 1/2
tSNE P 2/2
500
502
510

Figura 4B

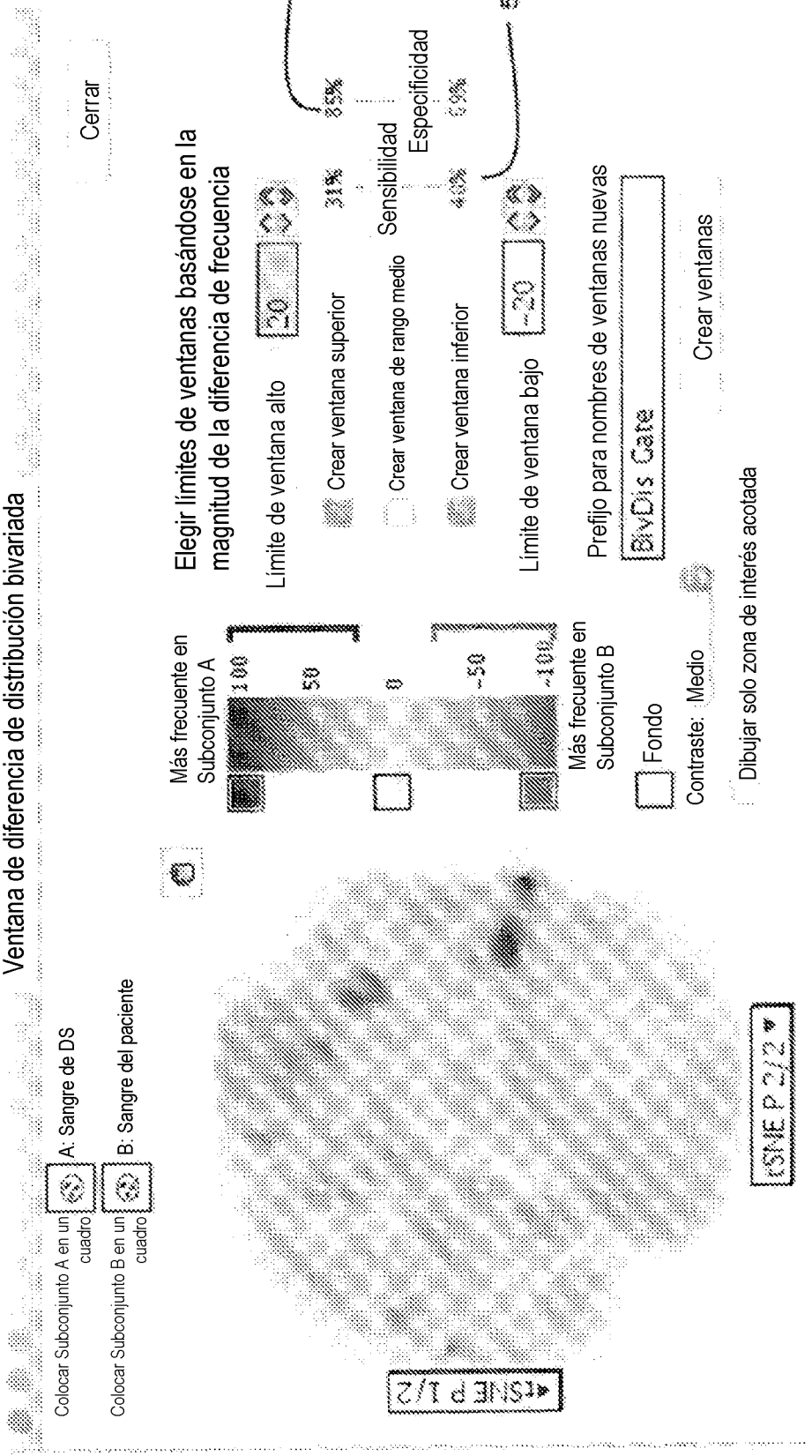
Ventana de diferencia de distribución bivariada
Cerrar
Colocar Subconjunto A en un cuadro
A: Sangre de DS
Colocar Subconjunto B en un cuadro
B: Sangre del paciente
Más frecuente en Subconjunto A
100
50
0
-50
-100
Más frecuente en Subconjunto B
Elegir límites de ventanas basándose en la magnitud de la diferencia de frecuencia
Límite de ventana alto
20
Crear ventana superior
31%
Crear ventana de rango medio
Crear ventana inferior
40%
Límite de ventana bajo
-20
Sensibilidad
Especificidad
522
520
Prefijo para nombres de ventanas nuevas
BivDis Cate
Fondo
Contraste: Medio
Crear ventanas
Dibujar solo zona de interés acotada

Figura 4C