(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 564 632**

(51) Int. CI.:

A61K 49/00 (2006.01)
G06F 19/00 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **26.03.2007 E 07754043 (3)**

(97) Fecha y número de publicación de la concesión europea: **17.02.2016 EP 2007434**

(54) Título: **Método y sistema para determinar si un fármaco será eficaz en un paciente con una enfermedad**

(30) Prioridad:

31.03.2006 US 396328

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
28.03.2016

(73) Titular/es:

BIODESIX INC. (100.0%)
2790 Wilderness Place, Suite 100
Boulder, CO 80301, US

(72) Inventor/es:

RÖDER, HEINRICH;
TSYPIN, MAXIM y
GRIGORIEVA, JULIA

(74) Agente/Representante:

ISERN JARA, Jorge

DESCRIPCIÓN

Método y sistema para determinar si un fármaco será eficaz en un paciente con una enfermedad

Antecedentes

Los inventores de la presente invención han encontrado un método novedoso para determinar si un paciente responderá a un tratamiento mediante el análisis de los biomarcadores del paciente mediante espectroscopia de masas. Como ejemplo de una realización de esta invención, los inventores han aplicado su técnica a un cáncer, el cáncer no microcítico de pulmón (CNMP).

El cáncer no microcítico de pulmón es una causa principal de muerte por cáncer tanto en hombres como en mujeres en los Estados Unidos. Existen al menos cuatro (4) tipos distintos de CNMP, que incluyen adenocarcinoma, carcinoma de células escamosas, de células grandes, y broncoalveolar. El carcinoma de células escamosas (epidermoide) de pulmón es un tipo de cáncer microscópico que se relaciona principalmente con el hábito tabáquico. El adenocarcinoma de pulmón es responsable de más del 50 % de todos los casos de cáncer de pulmón en EE.UU. Este cáncer es más frecuente en las mujeres y, además es el tipo más frecuente observado en los no fumadores. El carcinoma de células grandes, especialmente los que tienen características neuroendocrinas, se asocia habitualmente con la diseminación de tumores al cerebro. Cuando el CNMP entra en el torrente sanguíneo, puede diseminarse hasta sitios distantes tales como el hígado, los huesos, el cerebro, y a otros lugares del pulmón.

El tratamiento del CNMP ha sido relativamente pobre a lo largo de los años. La quimioterapia, el pilar de los tratamientos de los cánceres avanzados, es eficaz sólo ligeramente, con la excepción de los cánceres localizados. Aunque la cirugía es la opción potencialmente más curativa para el CNMP, no siempre es posible, dependiendo del estadio del cáncer.

Los planteamientos recientes para el desarrollo de fármacos anticancerosos para el tratamiento del paciente con CNMP se centran en reducir o eliminar la capacidad de las células cancerosas para crecer y dividirse. Estos fármacos anticancerosos se usan para alterar las señales que les dicen a las células si crecer o morir. Normalmente, el crecimiento celular está estrechamente controlado por las señales que reciben las células. En el cáncer, sin embargo, esta señalización va mal y las células siguen creciendo y dividiéndose de forma incontrolable, formando de esa forma un tumor. Una de estas vías de señalización comienza cuando una sustancia química del cuerpo, llamada factor de crecimiento epidérmico, se une a un receptor que se encuentra en la superficie de numerosas células del cuerpo. El receptor, conocido como el receptor del factor de crecimiento epidérmico (EGFR, de *epidermal growth factor receptor*) envía una señal a las células, a través de la activación de una enzima llamada tirosina cinasa (TK, de *tyrosine kinase*) que se encuentra dentro de las células. Las señales se usan para notificar a las células que crezcan y se dividan.

Dos fármacos anticancerosos que se desarrollaron y prescribieron para el CNMP se llaman gefitinib (nombre comercial "Iressa") y erlotinib (nombre comercial "Tarceva"). Estos fármacos anticancerosos se dirigen hacia la vía del EGFR y han mostrado una eficacia prometedora en el tratamiento del CNMP. Iressa inhibe la enzima tirosina cinasa, que está presente en las células cancerosas del pulmón, así como en otros cánceres, en tejidos normales, y que parece ser importante para el crecimiento de las células cancerosas. Iressa se ha usado como agente individual para el tratamiento del CNMP que ha avanzado después de, o que no ha respondido a, otros dos tipos de quimioterapias.

Sin embargo, los índices de respuesta han estado solo entre 10 % y 20 % en poblaciones de raza blanca, y han llevado a la Agencia Federal del Medicamento (FDA, de *Federal Drug Administration*) en 1995 a retirar su apoyo para la aplicación de Iressa como tratamiento de segunda línea. Sorprendentemente, el índice de respuesta en Asia ha sido considerablemente mayor e Iressa se sigue usando. Tarceva aún está aprobado y sigue administrándose de forma rutinaria a los pacientes, pero aún tiene índices de respuesta preocupantes. Aunque parece que Iressa y Tarceva tienen la capacidad de ser eficaces en algunos pacientes, pueden no ser fármacos eficaces de forma genérica al tratar a todos los pacientes. Puede haber muchos factores implicados en la capacidad de un paciente para responder a estos fármacos que actualmente se desconocen. Sin embargo, si pudiera usarse una determinación de factores que para predecir la eficacia de un paciente con CNMP para responder a estos fármacos anticancerosos, la FDA podría permitir que se prescribieran estos fármacos anticancerosos a aquellos pacientes que tuvieran las condiciones que indicasen que podrían responder a estos fármacos. Los doctores podrían entonces prescribir estos fármacos a los pacientes que se ha predicho que van a responder a los fármacos anticancerosos con el conocimiento de que sus pacientes podrían responder a los tratamientos.

El documento WO2005/098445 divulga métodos para la identificación de biomarcadores y de su uso en el diagnóstico del cáncer de pulmón. Los métodos diferencian entre individuos con cáncer de pulmón e individuos normales. El documento WO 2005/010492 divulga un método para la identificación de características biológicas mediante la recolección de un conjunto de datos que se relacionan con individuos que tienen características biológicas conocidas y el análisis del conjunto de datos para identificar biomarcadores que se relacionan potencialmente con clases de estado biológico seleccionadas. También se proporciona una metodología para la

utilización de datos de espectroscopia de masas para identificar biomarcadores peptídicos y protéicos que pueden usarse para discriminar de forma óptima entre muestras experimentales y de control - en el que las muestras experimentales pueden proceder, por ejemplo, de pacientes con cáncer de pulmón.

Sumario

La presente invención se relaciona con un método como se define mediante la reivindicación 1.

Para superar el problema de los bajos índices de éxito en el tratamiento usando fármacos, los principios de la presente invención proporcionan un análisis diagnóstico para determinar si un paciente responderá a estos tratamientos farmacológicos. La determinación se efectúa mediante la detección de picos de diferenciación de un espectro producido por un espectrómetro de masas a partir de suero extraído de la sangre de un paciente. Los biomarcadores son parámetros biológicos determinables y cuantificables que pueden evaluarse como un indicador de procesos biológicos normales o anómalos, o de procesos patogénicos. El espectrómetro de masas produce un espectro que tiene ciertos picos que pueden utilizarse para comparar con los espectros producidos a partir de suero de pacientes que respondieron o no respondieron a los tratamientos farmacológicos. No suele ser necesario determinar realmente qué compuesto químico está localizado en el pico. El propio espectro es una valiosa huella dactilar que puede caracterizar el tratamiento potencial para el fármaco en un paciente específico. El material que está en los picos puede aislarse y puede determinarse qué materiales están elevados o disminuidos en la muestra.

El documento WO 2005/010492 divulga un método para la identificación de características biológicas mediante la recolección de un conjunto de datos que se relacionan con individuos que tienen características biológicas conocidas y el análisis del conjunto de datos para identificar biomarcadores que se relacionan potencialmente con clases de estado biológico seleccionadas. También se proporciona una metodología para la utilización de datos de espectroscopia de masas para identificar biomarcadores peptídicos y protéicos que pueden usarse para discriminar de forma óptima entre muestras experimentales y de control - en el que las muestras experimentales pueden proceder, por ejemplo, de pacientes con cáncer de pulmón.

De forma más específica, los principios de la presente invención se dirigen a un proceso para la determinación de si un paciente con un cáncer no microcítico de pulmón responderá a un fármaco, en el que el fármaco es gefitinib o erlotinib. El proceso incluye la obtención de un espectro de análisis producido por un espectrómetro de masas a partir de un suero de un paciente. El espectro de análisis se procesa para determinar una relación con un grupo de espectros etiquetados por clase producidos a partir del suero respectivo de otros pacientes en el mismo o similar estadio clínico de la enfermedad o trastorno y que se sabe que han respondido o no han respondido al fármaco. Sobre la base de la relación del espectro de análisis con el grupo de espectros etiquetados por clase, puede efectuarse una determinación de si el paciente responderá al fármaco. Al procesar el espectro de análisis, puede realizarse reducción del fondo, normalización y alineamiento del espectro de análisis para emparejar mejor el espectro de análisis con el grupo de espectros etiquetados por clase, que se han procesado de igual o similar forma. Mediante el procesamiento de los espectros brutos para generar los espectros etiquetados por clase, puede efectuarse la determinación de si el fármaco será eficaz independientemente de la clínica particular y de los espectrómetros de masas usados para procesar el suero del paciente.

Otros ejemplos de acuerdo con los principios de la presente invención incluyen sistemas para determinar si un paciente responderá a un fármaco o tratamiento. Los sistemas pueden incluir un dispositivo de almacenamiento configurado para almacenar un espectro de análisis producido por un espectrómetro de masas a partir de un suero producido a partir de un paciente con una enfermedad o trastorno y un grupo de espectros etiquetados por clase producidos a partir del suero respectivo de otros pacientes en el mismo o similar estadio clínico de la enfermedad o trastorno y que se sabe que han respondido o no han respondido al fármaco o tratamiento. Dichos sistemas pueden incluir además un procesador en comunicación con un dispositivo de almacenamiento, en el que el procesador ejecuta el programa para (i) obtener un espectro de análisis producido por un espectrómetro de masas a partir de un suero producido a partir de paciente que tiene una enfermedad o trastorno, (ii) procesar el espectro de análisis para determinar una relación con un grupo de espectros etiquetados por clase producidos a partir del suero respectivo de otros pacientes que tienen el mismo o similar estadio clínico de la enfermedad o trastorno y que se sabe que han respondido o no han respondido a un fármaco o tratamiento. (iii) determinar, sobre la base de la relación del espectro de análisis con el grupo de espectros etiquetados por clase, si el paciente responderá al fármaco. En un ejemplo, el sistema está en comunicación con una red, tal como Internet, para la comunicación con los laboratorios y clínicas que comunican los espectros de análisis para analizar. La determinación de la relación de los espectros de análisis con el grupo de espectros etiquetados por clase puede incluir la producción de un indicador o representante de etiqueta de clase de respuesta potencial del paciente al fármaco o tratamiento. El indicador puede ser positivo, negativo, o no concluyente, de modo que un profesional médico pueda determinar si prescribir o no el fármaco o tratamiento. En algunos ejemplos, la enfermedad o trastorno es cáncer. En otros ejemplos, el tipo de cáncer es carcinoma no microcítico de pulmón. Aún en otro ejemplo, el sistema puede utilizarse para determinar si el fármaco gefitinib y/o erlotinib será eficaz en el tratamiento de los pacientes con carcinoma no microcítico de pulmón.

Breve descripción de los dibujos

La fig. 1 es un diagrama por bloques de una relación de ejemplo entre un centro procesador de análisis de laboratorio, clínicas de investigación del cáncer, y clínicas de pacientes con cáncer;
La fig. 2 es un diagrama por bloques de un sistema de ejemplo para la comunicación y procesamiento de la información entre el centro procesador de análisis de laboratorio, las clínicas de investigación del cáncer, y las clínicas de pacientes con cáncer de la fig. 1;
La fig. 3 es un diagrama de flujo de un proceso de flujo de trabajo de ejemplo para el desarrollo de un análisis para determinar si un paciente con cáncer responderá a un fármaco anticanceroso de acuerdo con los principios de la presente invención;
La fig. 4 es una imagen de un gráfico de gel de ejemplo de todos los espectros usados en el desarrollo de un análisis;
La fig. 5 es un histograma que muestra la producción de un conjunto de puntos de datos de ejemplo a partir de un espectrómetro que tiene componentes de fondo y de señal;
Las fig. 6A y 6B son gráficas muestran un espectro con fondo y sin fondo después de que el fondo se haya sustraído del espectro, respectivamente;
La fig. 7A es una gráfica que muestra múltiples espectros que han sido completamente preprocesados para simplificar la comparación de los espectros como se muestra en la fig. 7B;
Las figs. 8A y 8B son gráficas que muestran múltiples espectros de muestra que están alineados;
La fig. 9 es una gráfica de un proceso de ejemplo para la selección de una característica mediante la localización de un pico frecuente en más de x espectros que tienen una cierta anchura;
La fig. 10 es una gráfica representativa de los espectros promedio en los grupos clínicos EP, EP -precoz, RP, EE-corta, y EE larga promediados a lo largo de todas las muestras disponibles del desarrollo del análisis en sus grupos respectivos;
La fig. 11 es una gráfica que muestra un grupo de indicios de espectros etiquetados por clase de ejemplo representativos de dos clases diferentes de avance de la enfermedad e indicios de un espectro de análisis para clasificar;
La fig. 12 es una gráfica de Kaplan-Meier de los datos de análisis que muestra los índices de supervivencia de grupos de pacientes clasificados de acuerdo con los principios de la presente invención obtenidos a partir del uso de muestras italianas como conjunto de práctica y de muestras japonesas como conjunto de análisis;
La fig. 13 es una gráfica de Kaplan-Meier de los datos de análisis que muestra los índices de supervivencia de grupos de pacientes clasificados de acuerdo con los principios de la presente invención obtenidos a partir del uso de las muestras japonesas como conjunto de práctica y de las muestras italianas como conjunto de análisis;
La fig. 14 es una gráfica de Kaplan-Meier de los datos de análisis que muestra los índices de supervivencia de grupos de pacientes clasificados de acuerdo con los principios de la presente invención generados mediante un algoritmo clasificador para un conjunto de muestras con ocultación total; y
La fig. 15 es un diagrama por bloques de un proceso de ejemplo para determinar si un paciente con cáncer responderá a un fármaco anticanceroso de acuerdo con los principios de la presente invención.

Descripción detallada de los dibujos

La fig. 1 es un diagrama por bloques de una relación de ejemplo entre un centro procesador de análisis de laboratorio 102, las clínicas de investigación del cáncer 104a-104n (colectivamente 104), y las clínicas de pacientes con cáncer 106a-106m (colectivamente 106). El centro de procesador de análisis de laboratorio 102 opera para procesar los análisis de las clínicas de investigación del cáncer 104 y las clínicas de pacientes con cáncer 106. En un ejemplo, las clínicas de investigación del cáncer 104 y las clínicas de pacientes con cáncer 106 son parte de la misma organización, tal como un hospital. Las clínicas de investigación del cáncer 104 realizan ensayos farmacológicos y pruebas para determinar la eficacia de ciertos fármacos para tratar pacientes. Por ejemplo, los pacientes con carcinoma no microcítico de pulmón que han pasado por estudios clínicos y pruebas de diversos fármacos anticancerosos para controlar el crecimiento y la diseminación de las células cancerosas tienen diferentes respuestas frente a los fármacos anticancerosos. Estos fármacos anticancerosos pueden incluir gefitinib y erlotinib, que se dirigen hacia la vía del receptor del factor de crecimiento epidérmico. Durante los estudios clínicos y no clínicos, las clínicas de investigación del cáncer 104 controlan cuidadosamente diversos aspectos de los tratamientos, incluyendo estadios del cáncer, componentes sanguíneos, avance del cáncer, salud general del paciente, y otros factores indicativos del paciente para determinar la eficacia del fármaco anticanceroso.

La clínica de investigación del cáncer 106 puede ser cualquier centro que realice estudios clínicos o administre de otra forma medicamentos para el cáncer a pacientes con cáncer y controle la eficacia de los medicamentos. La clínica de investigación del cáncer 104 puede tomar muestras de sangre y procesarlas para producir suero, que es plasma sanguíneo (el componente líquido de la sangre en el que están suspendidas las células sanguíneas) con los factores de coagulación, tales como fibrina, eliminados. El suero puede procesarse y usarse para producir un espectro mediante un espectrómetro de masas de modo que puedan detectarse los biomarcadores del suero. En un ejemplo, el espectrómetro de masas es un espectrómetro de masas de tiempo de vuelo (TOF, de *time-of-flight*) que usa desorción/ionización láser asistida por matriz (MALDI, de *matrix-assisted laser desorption/ionization*). El espectro puede incluir marcadores o picos sustitutos dentro del espectro (véase la fig. 11) indicativos de ciertas sustancias químicas o materia dentro del suero.

Como resultado de la producción de espectros de los pacientes por el espectrómetro de masas, puede registrarse y observarse la eficacia de los fármacos anticancerosos que se están administrando al paciente con cáncer para producir resultados clínicos. El centro procesador de análisis de laboratorio 102 puede usar los resultados registrados (cuantitativos) y observados (de la salud general) de los pacientes para la determinación de clasificaciones para cada uno de los pacientes con cáncer en cuanto a si cada uno de ellos responde al (los) fármacos anticanceroso(s).

Continuando con la fig. 1, Como resultado de la producción de espectros de los pacientes por el espectrómetro de masas, puede registrarse y observarse la eficacia de los fármacos anticancerosos que se están administrando al paciente con cáncer para producir resultados clínicos. El centro procesador de análisis de laboratorio 102 recibe los espectros brutos con los resultados clínicos conocidos asociados 108 de las clínicas de investigación del cáncer y realiza una clasificación de cada espectro. La clasificación de cada espectro, descrita en detalle de aquí en adelante, clasifica cada espectro asociado con un paciente con cáncer que está recibiendo fármacos anticancerosos como con respuesta, sin respuesta, o con respuesta parcial. La clasificación de los espectros permite al centro procesador de análisis de laboratorio 102 recibir los espectros de análisis 110a-110m (colectivamente 110) de las clínicas de pacientes con cáncer 106 y realizar el análisis de estos espectros de análisis 110 para determinar a qué clasificación es más probable que se parezca cada espectro de análisis (es decir, de cada paciente). De forma alternativa, en vez de recibir los espectros brutos, el centro procesador de análisis de laboratorio 102 puede recibir muestras de sangre o muestras de suero para procesar y producir los espectros brutos para su procesamiento y clasificación.

Al clasificar el espectro bruto, se toma una decisión sobre si cada espectro es "bueno" o "malo" basada en si el paciente con cáncer tuvo una respuesta positiva, ninguna respuesta, o una respuesta limitada al fármaco anticanceroso. Mediante la comparación del espectro de análisis de los pacientes con cáncer con los espectros etiquetados por clase, puede hacerse una determinación en cuanto a la probabilidad de que un paciente con cáncer a partir del cual se ha generado un espectro de análisis tenga una respuesta positiva al fármaco anticanceroso. De aquí en adelante se proporciona una descripción más detallada del proceso de comparación. Una vez que el centro procesador de análisis de laboratorio 102 ha clasificado el espectro de análisis 110, y, de forma opcional, efectúa la determinación en cuanto a si el paciente con cáncer tendrá una respuesta positiva al fármaco anticanceroso, pueden entregarse los resultados de la clasificación 112a-112m (colectivamente 112) a la clínica de pacientes con cáncer respectiva 108a, por ejemplo. En una realización, los resultados de la clasificación son etiquetas de clase producidas por una función clasificadora como se describe posteriormente de aquí en adelante.

Aunque se muestra por separado, el centro de procesador de análisis de laboratorio 102 puede ser parte de las clínicas de investigación del cáncer 104 o las clínicas de pacientes con cáncer 106. En un ejemplo, el centro procesador de análisis de laboratorio 102 integra de forma funcional un equipo de análisis, tal como un espectrómetro de masas o un sistema de procesamiento que opera en conjunción con el equipo de análisis. De forma alternativa, la funcionalidad puede incorporarse a un sistema informático u otro sistema de procesamiento que esté configurado para realizar el procesamiento diverso utilizado en el procesamiento y clasificación de los espectros y no como parte o asociado al equipo de análisis. Por ejemplo, el sistema informático puede ser un servidor manejado por el centro procesador de análisis de laboratorio 102, la clínica de investigación del cáncer 104, y/o la clínica de pacientes con cáncer 106.

Aunque la fig. 1 describe las clínicas de cáncer, debe entenderse que estas clínicas pueden ser clínicas comunes o clínicas específicas de un enfermedad o afección concreta. Por consiguiente, el centro procesador de análisis de laboratorio 102 está configurado para recibir y analizar la enfermedad o afección particular que se está enviando de acuerdo con los principios de la presente invención.

La fig. 2 es un diagrama por bloques de un sistema de ejemplo 200 para la comunicación y procesamiento de la información entre el centro procesador de análisis de laboratorio 102, las clínicas de investigación del cáncer 104, y las clínicas de pacientes con cáncer 106 de la fig. 1. Un centro procesador de análisis de laboratorio 102 puede manejar un sistema informático 200 de un centro procesador de análisis de laboratorio. Las clínicas de investigación del cáncer 104 pueden manejar los servidores 204a-204n (colectivamente 204) de las clínicas de investigación del cáncer y las clínicas de pacientes con cáncer 106 pueden manejar los servidores 206a-206m de las clínicas de pacientes con cáncer. Cada uno de los sistemas informáticos 202 y de los servidores 204 y 206 puede comunicarse a lo largo de la red 208 por medio de los paquetes de datos digitales 209a-209b u otra técnica de comunicación tal como se entiende en la técnica. La red 208 puede ser Internet u otra red de comunicación pública o privada.

El sistema informático 202 del centro procesador de análisis de laboratorio puede incluir un programa informático 212 que ejecute el procesador 210 para el procesamiento de los espectros brutos y los espectros de análisis para determinar las clasificaciones de todos o de una porción de los mismos de acuerdo con los principios de la presente invención tal como se describe posteriormente de aquí en adelante. El sistema informático 202 puede incluir además la memoria 214, en la cual puede residir el programa 212 cuando se esté ejecutando, la unidad de entrada/salida (I/O) (de *input/output*) 216, la cual puede realizar la comunicación a lo largo de la red 208, y el dispositivo de almacenamiento 218 con el cual comunica el procesador. El dispositivo de almacenamiento 218 puede incluir una o más bases de datos 220a-220n (colectivamente 220) en las que se almacenan los espectros de análisis, y otros datos relacionados para permitir que el centro procesador de análisis de laboratorio 102 determine si un paciente

con cáncer responderá a un fármaco anticanceroso. Debe entenderse que el dispositivo de almacenamiento 218 puede incluir uno o más dispositivos de almacenamiento y situarse dentro o externamente al sistema informático 202. Debe entenderse además que el procesador 210 puede incluir uno o más procesadores. Más aún, debe entenderse que el sistema informático 202 puede estar directa o indirectamente en comunicación con la red 208.

De acuerdo con la fig. 1, los servidores 204 de la clínica de investigación del cáncer pueden comunicar los espectros brutos con los resultados clínicos conocidos asociados basados en los ensayos clínicos del fármaco anticanceroso al sistema informático 202 del procesador de análisis de laboratorio. El procesador 210, puede realizar el procesamiento de la clasificación para clasificar cada espectro bruto para clasificar los espectros brutos para formar un grupo de espectros clasificados de forma automática o semiautomática con la asistencia de un científico o de otra forma. De forma similar, los servidores 206 de la clínica de pacientes con cáncer pueden comunicar los espectros de análisis 110 al laboratorio para que el procesador 210 clasifique de forma automática o semiautomática los espectros de análisis 110 para las clínicas de pacientes con cáncer 108. El sistema informático 202 del centro procesador de análisis de laboratorio puede procesar los espectros de análisis 110 y comunicar como respuesta los resultados de la clasificación 112 (fig. 1) a los servidores 206 de la clínica de pacientes con cáncer. Como resultado de la clasificación de los espectros brutos y los espectros de análisis 112, el sistema informático 202 puede almacenar los resultados de la clasificación y utilizar los resultados para generar información estadística que puede usarse para otros fines diversos, tales como los índices de éxito y fracaso del fármaco anticanceroso.

El análisis de datos juega un papel crucial en el descubrimiento de los picos de detección que diferencian a los espectros de pacientes con diferente resultado clínico y de su uso indistinto como ejemplos de descubrimiento para ensayos de inmunohistoquímica o directamente en diagnóstico basado en espectrometría de masas. Al desarrollar los procedimientos de ensayo y análisis de acuerdo con los principios de la presente invención, se ha desarrollado un sistema de análisis integrado que contiene algoritmos diseñados para el análisis comparativo de los espectros de masas. El sistema de análisis integrado incluye un número de herramientas que facilita la detección de los picos de diferenciación en los espectros de los espectros de masas, mientras que, al mismo tiempo proporciona herramientas rigurosas para la evaluación de su significación y la validación de los resultados.

La fig. 3 es un diagrama de flujo de un proceso de flujo de trabajo de ejemplo 300 para el desarrollo y la realización de un análisis para determinar si un paciente con cáncer responderá a un fármaco anticanceroso de acuerdo con los principios de la presente invención. El proceso empieza en la etapa 302 en la que se toman muestras de los pacientes con cáncer. Dependiendo del tipo de cáncer o de otra enfermedad, pueden utilizarse rociados de tejido, lisados celulares, o células cortadas como muestras para la generación de espectros por medio de un espectrómetro de masas 304. El espectrómetro de masas puede ser un espectrómetro de masas ABI Voyager, un ABI 4700, un Bruker Autoflex o un espectrómetro de masas Bruker Ultraflex. Pueden utilizarse otros espectrómetros de masas de forma similar. En el caso del carcinoma no microcítico de pulmón, puede usarse suero para generar los espectros. Mediante el uso de suero, los pacientes de cáncer de pulmón en estadios avanzados del cáncer de pulmón, en los que es difícil o imposible tomar una muestra de tejido de los pacientes, pueden diagnosticarse sin un procedimiento invasivo. De forma adicional, pueden utilizarse líquidos corporales, tales como la orina, para muestras en la detección de picos en un espectro de masas para determinar si ciertos fármacos anticancerosos serán eficaces en el tratamiento de un paciente con cáncer con carcinoma no microcítico de pulmón. Usando procedimientos no invasivos para la recolección de suero u otros fluidos, el coste del diagnóstico es significativamente menor que si se necesitara una muestra de tejido de un pulmón.

La generación y el procesamiento del suero usado para un estudio del análisis pueden incluir el uso de muestras de suero no procesado de hospitales individuales. En un ejemplo, las muestras de suero no procesado pueden descongelarse en hielo y centrifugarse a 1500 rpm durante cinco minutos a cuatro grados Celsius. Después, las muestras de suero pueden diluirse a 1:10, como se realiza en la el Centro de Ciencias de la Salud de la universidad de Colorado (UCHSC, *University of Colorado Health Sciences*) o a 1:5, como se realiza en el Centro Médico de la Universidad de Vanderbilt (VUMC, *Vanderbilt University medical Center*), en agua MilliQ. Las muestras diluidas pueden rociarse en posiciones situadas al azar sobre una placa MALDI por triplicado (es decir, en tres dianas MALDI diferentes). Tras rociar 0,75 µl de suero diluido sobre una placa MALDI, pueden añadirse 0,75 µl de 35 mg/ml de ácido sinapínico (en acetonitrilio al 50 % y TFA al 0,1 %) y mezclarse mediante pipeteo hacia arriba y hacia abajo cinco veces. Las placas se pueden dejar secar a temperatura ambiente. Debe entenderse que pueden utilizarse otras técnicas y procedimientos para la preparación y el procesamiento del suero de acuerdo con los principios de la presente invención.

Los espectros de masas pueden adquirirse para iones positivos en modo lineal usando un Voyager DE-PRO (UCHSC) o DE-STR (VUMC) con recolección de los espectros automatizada o manual. En un estudio, se recogieron 75 (UCHSC) o 100 (VUMC) espectros a partir de siete (UCHSC) o cinco (VUMC) posiciones dentro de cada rociado de MALDI a fin de generar un promedio de 525 (UCHSC) o 500 (VUMC) espectros para cada muestra de suero. Los espectros se calibraron externamente usando una mezcla de patrones de proteína (insulina (bovina), tiorredoxina (*E. coli*), y apomioglobina (equina)). Con fines de validación, se llevaron a cabo tres réplicas de la misma muestra para todas las muestras, dando como resultado 717 espectros (239 muestras, 3 veces) presentados para su análisis para el presente estudio.

Al realizar el análisis de los datos, se acepta, en general, que las células cancerosas tienen diferente nivel de expresión de proteínas específicas que son diferentes de las células normales. Estadios diferentes de enfermedad van acompañados por cambios en proteínas específicas, p. ej., cambios en el nivel de expresión de proteínas de unión en el caso del cáncer metastásico. En el caso de las muestras de suero, y para definir el análisis del suero del análisis de muestras de tejido, es improbable que se determinen excreciones directas del tumor debido a la dilución de estas excreciones en la sangre. Los picos de diferenciación en las muestras de suero (o de otros líquidos corporales) surgen con toda probabilidad debidos a una reacción de respuesta del anfitrión dependiente del estadio de la enfermedad, tal como las reacciones autoinmunes. Por definición, hay que esperar que los análisis basados en muestras de tejidos sean altamente específicos, pero no necesariamente muy significativos, y los análisis con espectrómetro de masas basados en suero deberían ser altamente significativos, pero no tan específicos. Esto nace de los resultados presentados de aquí en adelante. Mediante la detección de picos de diferenciación en los espectros, puede realizarse una correlación de los cambios con cuestiones clínicamente relevantes. Para generar picos de diferenciación en los espectros de valor, independientemente de su uso posterior, bien directamente, como herramienta diagnóstica, o bien como ejemplos para el análisis basado en la inmunohistoquímica, pueden abordarse las siguientes cuestiones durante el proceso de descubrimiento de los picos de diferenciación, incluyendo la etapa de análisis de los datos:

Reproducibilidad: los resultados de un análisis han de ser reproducibles. Pueden identificarse biomarcadores a través de los picos de diferenciación que pueden encontrarse de forma repetitiva en los diversos grupos de enfermedad y de control, y los valores asignados a estos picos de diferenciación no pueden variar demasiado dentro de un grupo. Como medida simplificada de reproducibilidad, el programa ejecutado en un procesador puede proporcionar coeficientes de variaciones (CV), que se han convertido en un patrón para evaluar pruebas diagnósticas. Las variaciones de los marcadores dentro de un grupo, e incluso dentro de la misma muestra, pueden determinarse, caracterizarse, y usarse en el análisis y clasificación ulterior.

Consistencia: Los picos de diferenciación han de ser consistentes frente a las variaciones inevitables en la preparación y manipulación de la muestra, así como frente a las variaciones surgidas de tendencias en las características del espectrómetro de masas. Otra razón para la variabilidad de paciente a paciente surge de diferencias irrelevantes en el estado biológico del paciente, por ejemplo, el estado digestivo en el momento de la recogida de la muestra. Pueden desarrollarse criterios para distinguir los cambios irrelevantes de los biológicamente significativos. Al diseñar los clasificadores (es decir, funciones o algoritmos clasificadores), que son funciones que cartografían desde el espacio multidimensional de características (p. ej., 12 picos de diferenciación) hasta el espacio de la etiqueta por clase (p. ej., "bueno", "malo" o "no definido") y durante la extracción de características, los picos reales de diferenciación no deberían cambiar mucho al hacer pequeños cambios en los parámetros de análisis de los datos. Los picos de diferenciación situados de forma similar deberían encontrarse en conjuntos de datos diferentes.

Interpretabilidad: Los picos de diferenciación resultantes pueden ponerse en el contexto de la interpretabilidad biológica. En primer lugar, los picos de diferenciación identificados son generalmente evidentes de forma visual en los espectros de masas. Las posiciones m/z de los picos de diferenciación dan información valiosa sobre la importancia biológica de los marcadores subyacentes que causan estos picos de diferenciación. Esto permite la interpretación y el filtrado de los picos de diferenciación que surgen a partir de procedimientos biológicamente irrelevantes Por ejemplo, las determinaciones del diferente contenido en hemoglobina de las muestras cancerosas frente a las normales, el cual es puramente un artefacto de la preparación de la muestra. En algunos casos, puede resultar que picos de diferenciación del espectro clínicamente relevantes sean combinaciones no lineales de múltiples características del espectro, y no sean simples regulaciones positivas y negativas. Incluso en este caso, los picos de diferenciación que constituyen características en los espectros deberían ser visibles (fig. 4), y las funciones con las cuales evaluar los marcadores deberían hacerse explícitas.

Sensibilidad: Suele realizarse un gran esfuerzo para recoger muestras y generar espectros de masas. También se tiene un gran cuidado para evitar la pérdida de picos de diferenciación relevantes en los espectros del espectrómetro de masas mediante el uso de algoritmos de análisis de datos que no son selectivos o suficientemente sensibles para encontrar realmente estos picos de diferenciación en un espectro. Por ejemplo, si se define un intervalo de m/z como relevante para una característica, este intervalo debe ser suficientemente grande para contener la característica, y no debería combinar otras características presentes en el espectro. Los algoritmos para la obtención del intervalo obtienen sus parámetros a partir de los propios datos, opcionalmente, de una manera local, y pueden no depender de parámetros externos de suavizado y fijación.

La tarea de comparar espectros de masas para la extracción de picos de diferenciación se hace difícil por la naturaleza específica de estos espectros debido a variaciones intrínsecas de la intensidad. La probabilidad de ionización de iones individuales depende de la química de la muestra local (p. ej., los efectos de supresión de iones), y aunque la resolución de masas del módem de los espectrómetros de masas suele ser suficiente, la escala absoluta de masa puede variar de espectro a espectro.

De acuerdo con los principios de la presente invención, las variaciones específicas del espectrómetro de masas pueden determinarse para reducir o eliminar estas variaciones (en caso de variaciones del fondo) o proporcionar

determinaciones para evaluar la significación relevante de las señales mediante la estimación del nivel de ruido local. Puede lograrse evitar la introducción de variaciones adicionales que surgen del preprocesamiento y el análisis de los datos. Por ejemplo, se ha encontrado que el programa de recogida de picos que suele ir en el lote junto con muchos espectrómetros no es fiable para usar directamente estos picos en un análisis espectral comparativo. Los primeros intentos en cuanto a la comparación espectral han recurrido, en cambio, al uso del total mismo de los espectros de masas para sus algoritmos de comparación y clasificación. El total de los espectros, sin embargo, incluye muchos miles de puntos de datos individuales, la mayoría de los cuales son determinaciones del ruido del equipo cuya única información relevante se limita a los picos de los espectros de masas. Además, la interpretación de las características de los espectros es complicada y a veces no lineal en el caso de los algoritmos de clasificación basados en redes neurales, y se hace muy engorrosa. Como resultado, la aplicación de estos intentos para clasificar las muestras de suero ha conducido a reivindicaciones exageradas que no pudieron reproducirse en otros laboratorios.

La fig. 4 es una imagen de un gráfico de gel de ejemplo 404 de los resultados de un espectrofotómetro para un marcador. El espectro se etiqueta clínicamente mediante el uso de patrones de etiquetas de progresión de la Organización Mundial de la Salud (OMS), que incluyen enfermedad estable (EE), enfermedad progresiva (EP), y pacientes con respuesta parcial (RP). Sin embargo, se crean etiquetas clínicas finas, que separan las etiquetas clínicas principales en etiquetas clínicas extremas, para incluir las tres etiquetas adicionales de EE-corta, EE-larga, y EP-precoz. Un gráfico de gel es un gráfico en el que cada línea corresponde a un espectro de masas de una muestra clínica, el eje horizontal es el eje de masa/carga, y la escala de grises ilustra la intensidad. Las etiquetas clínicas 402 se proporcionan en el gráfico de gel 404 con las líneas horizontales 404 delineando las diferentes etiquetas clínicas. El gráfico de gel 404 es el de todos los espectros (es decir, los espectros recibidos de una clínica de investigación del cáncer de un grupo de control de pacientes con cáncer no microcítico de pulmón en Italia y Japón que recibieron Iressa como tratamiento para el cáncer) usados para practicar un algoritmo clasificador. Los picos de diferenciación pueden verse de forma visual en cada uno de los espectros en 406 y 408, pero se determinan de forma cuantitativa a efectos de precisión y de otros fines cuantitativos.

Al evitar algunos de estos problemas de determinación, pueden preprocesarse los espectros de masas brutos para eliminar y determinar los artefactos irrelevantes del proceso de espectrometría de masas, y para registrarlos en una escala similar de m/z y amplitud.

Continuando con la fig. 3, el proceso en la etapa 306 realiza el preprocesamiento de datos. El preprocesamiento puede incluir cualquier o toda la sustracción del fondo, estimación del ruido, normalización, recogida de picos, y alineamiento espectral. Estos procesos se ilustran en las fig. 5-10 y se describen de aquí en adelante.

La fig. 5 es un histograma 500 que muestra un conjunto resultados de puntos de datos de ejemplo a partir de un espectrómetro que tiene componentes de fondo y de señal. El fondo o el valor basal es un componente de un espectro de masas que varía lentamente -el cambio global gradual de los datos a lo largo del intervalo m/z. Como definiciones funcionales: El fondo son las variaciones suaves de la fuerza de la señal que pueden surgir a partir de los efectos de la acumulación de carga o las características no lineales del detector o la degradación iónica parcial, etc., al contrario que el ruido, que surge a partir del sistema electrónico, los iones aleatorios, y fluctúa rápidamente (en m/z).

El fondo puede modelarse, y, por tanto, sustraerse. El fondo es una fluctuación estadística y sólo puede determinarse su fuerza. Además, el fondo puede estar causado por iones "basura" no resueltos y puede estimarse y sustraerse antes de que las etapas posteriores de procesamiento de los datos, tales como la detección de picos, puedan realizarse de forma significativa. El fondo puede estimarse usando estimadores estadísticos locales consistentes. La obtención de una estimación fiable para la fuerza del ruido en los datos se utiliza para la posterior detección de los picos basada en el criterio de la relación entre señal y ruido (S/R). Dichos estimadores también se usan en cualquier tarea de comparación espectral para proporcionar una determinación de los errores. Al igual que en la estimación del fondo, pueden utilizarse estimadores asimétricos consistentes para realizar esta tarea.

El fondo se muestra para incluir la mayor parte de los números de puntos de datos y la señal incluye menos puntos de datos. El fondo puede determinarse mediante la iteración usando un análisis de correlación y separación óptima. Como el fondo no contiene información biológicamente relevante y varía de espectro a espectro, la información de la amplitud puede hacerse más comparable mediante la sustracción del valor del fondo de cada espectro. Este proceso se describe en la publicación de la patente en trámite junto con la presente, con número de serie US 2005/0267689 presentada el 7 de julio de 2004.

Las fig. 6A y 6B son las gráficas 600a y 600b que muestran un espectro con fondo 602 y sin fondo después de haber sustraído el fondo del espectro 604, respectivamente. Como suele ocurrir en el suero, hay picos que son muy variables debido a fluctuaciones naturales en la abundancia de proteoma sérico. Además, la cantidad de muestra ionizada puede fluctuar de espectro a espectro debido a cambios en la potencia del láser, variaciones en la cantidad de muestra ionizable, y variaciones en la colocación del láser en la placa de MALDI. Esta fluctuación produce rutinas de normalización convencionales, tales como la normalización de la corriente iónica total (es decir, la normalización a lo largo de todo el espectro), menos útiles, ya que las fluctuaciones en estos picos se propagan hacia los picos de

interés. Puede utilizarse una normalización parcial (es decir, normalización a lo largo de los espectros que identifica y excluye estos picos y regiones variables), para evitar los resultados que fluctúan, proporcionando de esta forma resultados reproducibles.

Más concretamente, la normalización de la corriente iónica parcial puede obtenerse de la forma siguiente. El espectro de masas incluye puntos de datos, pares (m/z, amplitud), dispuestos en orden ascendente de m/z. Como el espectro se obtiene en un equipo de tiempo de vuelo, el eje m/z puede considerarse segmentado en intervalos. Cada punto de datos representa el intervalo correspondiente y su amplitud representa (es proporcional a) el recuento de iones del intervalo (es decir, la corriente iónica en el intervalo).

La suma de todas las amplitudes del espectro es, por tanto, la "corriente iónica total" (CIT). Corresponde al número total de iones que llegan a un detector del espectrómetro de masas. La normalización de la corriente iónica total significa que, para cada espectro, se elige un factor de normalización tal que los espectros normalizados correspondientes (m/z= m/z original, amplitud= (factor de normalización)*(amplitud original)) tengan la misma (prescrita) corriente iónica total, tal como 100.

En general, la normalización de la corriente iónica total sólo tiene sentido después de la sustracción del fondo. De otra forma, la corriente iónica total está dominada por el fondo integrado, en vez de por la corriente iónica en las señales significativas, tales como los picos. En otras palabras, la corriente iónica total integra todos los iones disponibles y está dominado por los picos grandes. En el caso en el que los picos son muy variables, la corriente iónica total también es muy variable, causando de esta forma variación de la normalización, la cual puede conducir a la detección falsa positiva de características de diferenciación.

De acuerdo con los principios de la presente invención, detección de "características" - intervalos del eje m/z que parecen estar "no vacíos", es decir, no de "fondo puro" debido a que contienen algo de señal, tales como picos. Una característica es un pico que es visible en más de un número de espectros definido por el usuario de un grupo de control de pacientes. Tener un conjunto de características (una colección de intervalos de m/z no solapantes) apoya la definición de un método de normalización más flexible, la "normalización de la corriente iónica parcial (CIP)". La corriente iónica parcial es la suma de las amplitudes del espectro para todos los puntos de datos que pertenecen al conjunto de características especificado (típicamente, un subconjunto del conjunto de características total). La normalización de la corriente iónica parcial significa que, para cada espectro, puede elegirse un factor de normalización tal que los espectros normalizados correspondientes (m/z= m/z original, amplitud= (factor de normalización)*(amplitud original)) tengan la misma (prescrita) corriente iónica parcial. En general, la corriente iónica parcial usa picos estables para la normalización, ya que los muy variables no se incluyen en los cálculos. Mediante el uso de picos estables, se obtiene como resultado la estabilidad en el proceso de normalización.

Los picos de los espectros dentro de un grupo control de pacientes se incluyen en una lista, y puede usarse un algoritmo de agrupamiento divisivo, tal como se entiende en la técnica, para encontrar grupos de picos.

TABLA I. Características de la normalización de la CIP

ID	m/z central	ID	m/z central	ID	m/z central	ID	m/z central
0	3085,867	33	9157,859	70	17168,815	97	28102,608
1	3102,439	34	9371,936	71	17272,049	98	28535,715
2	3107,451	35	9424,089	72	17391,032	99	28889,368
3	3129,212	36	9432,518	73	17412,315	100	28896,086
4	4154,918	37	9446,061	74	17590,691	101	28902,778
6	4187,865	38	9635,796	75	17620,442	102	33277,541
6	4711,48	39	9638,7	76	18629,158	103	33340,741
7	5104,862	40	9659,863	77	18824,353	104	33839,223
10	6433,973	41	9717,098	78	19104,212	105	38830,258
11	6588,426	42	9738,03	79	19460,971	106	43474,948
12	6591,603	43	9941,016	80	20868,776	107	50722,939
13	6632,237	44	10220,11	81	21040,264	108	56307,899
14	6839,537	45	10504,31	82	21063,912	109	57257,535
15	6883,021	46	10841,693	83	21275,194	110	59321,131

ID	m/z central	ID	m/z central	ID	m/z central	ID	m/z central
16	6941,514	53	12579,848	84	22690,405	111	65392,98
17	7390,573	54	12771,505	85	22844,388	112	66702,45
20	7673,52	55	12861,925	86	22927,864	113	67633,769
22	8206,572	56	12868,575	88	23215,972	114	68328,628
23	8230,679	57	13082,443	89	23354,353	115	73363,308
24	8697,822	58	13765,804	90	23451,251	116	77948,338
27	8822,777	59	13885,668	91	24917,423	117	91016,846
28	8880,021	60	14050,987	92	25147,019	118	96444,862
29	8920,239	61	14157,312	93	25185,861	119	98722,464
30	8940,18	62	14651,73	94	25466,131		
31	9135,182	68	16206,683	95	25582,933		
32	9138,189	69	17143,885	96	25813,574		

La tabla I incluye una lista del 80 % (CIP= 0,8) de todas las características (conjunto de características restante) que se conservaron en una normalización de la CIP. Los valores de m/z están en Daltons con una incertidumbre de 1000 ppm (después del alineamiento).

Un caso extremo de normalización de la corriente iónica parcial es cuando se usa el conjunto total de características para computar la corriente iónica parcial. Este caso es análogo a la normalización de la corriente iónica total, siendo la diferencia que las regiones "vacías" del espectro contribuyen a la corriente iónica total, pero no a la corriente iónica parcial. Por tanto, la contribución del ruido en la región "vacía" no se incluye en la corriente iónica parcial, Otro caso extremo es cuando sólo se usa una característica para computar la corriente iónica parcial. Si ésta es la característica que contiene el pico más fuerte, se determina la normalización del pico base.

En la comparación del espectro, el razonamiento detrás del uso de la normalización de la corriente iónica parcial es el siguiente. Considérense dos grupos de espectros, tales como de enfermedad y de control. Los espectros contienen del orden de 100 señales (picos), y se espera que la mayoría de las señales no cambien entre grupos, mientras que algunas señales pueden regularse positiva o negativamente. En los espectros de masas, las intensidades sin normalizar no son directamente comparables entre espectros. Cuando se usa la normalización de la corriente iónica total, se hace la suposición de que las señales reguladas positiva y negativamente son pocas y débiles, de modo que no distorsionan significativamente la corriente iónica total, la cual domina supuestamente las señales que no cambian entre grupos. Sin embargo, en realidad, éste no es necesariamente el caso. Si, por ejemplo, la señal regulada positivamente es lo suficientemente fuerte para contribuir de forma significativa a la corriente iónica total, otras señales de los datos normalizados aparecen como reguladas negativamente, aun cuando no hayan cambiado realmente. De forma análoga, si los espectros contienen señales fuertes y fuertemente variables, otras señales del espectro normalizado muestran coeficientes de variación aumentados, aun cuando sean estables por naturaleza. El uso de la normalización de la corriente iónica parcial en vez de la corriente iónica total, y el uso del subconjunto de características que contiene las características más estables, al tiempo que se omiten las características reguladas positivamente, reguladas negativamente o muy variables, puede solucionar el problema de los coeficientes de variación aumentados. La cuestión principal es cómo seleccionar este subconjunto.

Para seleccionar el subconjunto para la corriente iónica parcial, puede usarse el siguiente procedimiento. Si se obtienen varios grupos de espectros, para los fines de este procedimiento, los grupos de espectros pueden combinarse en un conjunto combinado.

En primer lugar, el conjunto de características equivale a una lista completa de características. A continuación, el siguiente procedimiento puede repetirse un número de veces para producir el nuevo subconjunto de las características "menos variables" que contienen una característica menos que en el original.

El proceso puede continuarse de la forma siguiente:

- Usando el subconjunto de características original, normalizar todos los valores de las características (conjunto completo) para la corriente iónica parcial.
- Para cada característica, computar el coeficiente de variación= (desviación típica)/(valor medio).
- Ordenar las características de acuerdo con el valor absoluto del CV.

- Seleccionar un nuevo subconjunto de características de entre esta lista ordenada -las de menor CV (abs; incluir una característica menos que en el subconjunto original.
- Sustituir el subconjunto original por el nuevo subconjunto.

Los criterios de terminación son los siguientes. El usuario especifica dos valores:

- la fracción mínima permitida de la corriente iónica
- la fracción mínima permitida del número de características.

El proceso se termina cuando se rompe alguno de los criterios. Por tanto, si el usuario especifica ambos valores (es decir, la fracción mínima permitida de la corriente iónica y del número de características) como 0,8, se garantiza que el subconjunto de características resultante contiene al menos el 80 % de la corriente iónica (computada a partir del conjunto de características completo), así como al menos el 80 % de las características. La especificación de 1,0 para cualquiera de los valores da como resultado el conjunto de características completo que se está usando. Típicamente, 0,8 está alrededor del valor correcto que hay que usar para unos resultados óptimos. Dependiendo de la aplicación, sin embargo, pueden usarse valores mayores o menores. Los valores de característica normalizados para la corriente iónica parcial pueden usarse entonces para clasificación u otros fines.

En resumen, la normalización de la corriente iónica parcial puede determinarse de la forma siguiente:

- calcular los CV
- descartar el pico con el mayor CV
- parar cuando el CV máximo sea menor de un nivel especificado.

La ejecución de la corriente iónica parcial puede calcularse usando dos operaciones. La primera operación computa una lista de características para su uso en el denominador de la CIP. Este marcador de la operación mezcla primero los valores de las características de los dos grupos seleccionados en una matriz bidimensional, en la que las filas son los espectros (es decir, las muestras) y las columnas son los valores de las características que corresponden en orden a la lista de características ordenada mediante CenterMZ. Esta operación toma dos parámetros además de los valores de características mezclados. Estos dos parámetros son MinAllowedFracOfIC y MinAllowedFracOfFeatures. MinAllowedFracOfIC - fracción mínima permitida de la corriente iónica en el subconjunto de características conservado. La conservación de estas características corresponde al valor 1. MinAllowedFracOfFeatures - fracción mínima permitida de características en el subconjunto de características conservado. La conservación de estas características corresponde al valor 1. Esta operación produce una ArrayList de números enteros, que representa los índices de las características que se van a usar en el denominador.

Un ejemplo de un algoritmo usado para llegar a la lista de características que usan la normalización de la CIP se resume en el siguiente pseudocódigo:

```
int n_samples = número de espectros en los dos grupos seleccionados;
int n_features <= número de características en la lista de características;

//construir la lista de todas las características
ArrayList NFList = new ArrayList ();
for (int j •= O; j < n_features; j++)
 {
      NFList. Add (j) }
}

//resultados en 1, como ésta es la lista de características completa está en NFList
Double frac_ic = FraclonCurrent(f, NFList);
//también da como resultado 1.
Double frac_f = ((double)NFList. Count) /n_features;
ArrayList NFList_old <= (ArrayList)NFList. Clone ();

//mientras que la fracción de la corriente IÓNICA sea mayor o igual
//a la especificada por el usuario, y el porcentaje de
//características usadas sea mayor que el especificado por el usuario.

While (frac_ic >= MinAllowedFracOfIC &&
          frac_f >= MinAllowedFracOfFeatures)
{
     NFList_old = (ArrayList)NFList.Clone ();
     //renormalizar sobre la base de la presente NFList, después computar
     //el coeficiente de variación para cada conjunto de
     //características normalizadas, después ordenar por coeficiente de variación,
```

```
        //el coeficiente de variación más alto se elimina de la
        //NFList.
        OneStep(f, ref NFList);
        //ahora hay una característica menos en la lista de índices,
        //y la fracción de la corriente IÓNICA puede computarse como el resultado de
        //CIP / CIT
        //en el que CIP es la suma de todos los valores de características para los
        //índices de características especificados en la NFList. Y donde
        //CIT es la suma de todos los valores de características.
        Frac_ic = FraclonCurrent (f, NFList);
        //frac f es simplemente el porcentaje de espectros que se están usando ahora
        //en la NFList
        Frac_f = ((double)NFList.Count)/n features;

                                                }
                                        return NFList old;
```

Numerosas variaciones menores y mayores adicionales en este algoritmo serán aparentes para un experto en esta técnica y se consideran como parte de la invención reivindicada.

Una vez que se ha completado el cálculo, se determina la lista de características que se va a usar en el denominador de la corriente iónica parcial.

La segunda operación es renormalizar todos los valores de características para los grupos especificados usando el denominador de la corriente iónica parcial. Se llega a los primeros valores de normalización para cada espectro/muestra usando los valores de características especificados por la lista de índices de resultados a partir de la operación previa. Después, estos valores de normalización se usan para modificar la lista de valores de características especificados dentro de la matriz bidimensional de valores de características.

Esta función se logra mediante la realización de un algoritmo representado mediante el siguiente pseudocódigo.

```
   Int n_samples = número de espectros en los dos grupos seleccionados;
   Int n_features <= número de características en la lista de características;

   //iniciar la matriz de resultado.
   Double[,] f2 «=» new double [n_samples, h_features];

   //matriz para los valores de normalización
   Double[] norm = new double[n_samples];

   //encontrar factores de normalización para cada muestra
   for (int I = 0; I < n_samples; I++)
   {
        norm[I] = 0;
        foreach (int k in NFList)
        {
           //ajustar norm[I] a la suma de todos los valores de características para los
           //índices de características especificados en la NFList.
           Norm[I].+= f[I,k];

        //dividir por el número de características especificadas.
        Normll] /= NFList.Count;

        for (int j = 0; J < n_ features; j++)
        {
           //normalizar dividiendo los valores de las características por el valor
           //de normalización para la característica especificada.
           F2[I,j] - f(I,j]/norm[I];
       }
   }
//devolver resultado
Return f2;
```

Numerosas variaciones menores y mayores adicionales en este algoritmo serán aparentes para un experto en esta técnica y se consideran como parte de la invención reivindicada.

Tras completar estas dos etapas, se completa la normalización de la corriente iónica parcial. La normalización de la corriente iónica parcial puede conducir a una reducción bastante drástica de los CV de los picos individuales. Para la reproducibilidad de los datos en orina, en la que se determina la variabilidad del preprocesamiento de la muestra por medio de fraccionamiento (resina para eliminar las sales), la reducción del CV es de alrededor de un factor de dos.

La fig. 7A es una gráfica 700a que muestra los espectros múltiples 702 y 704 que se han normalizado para simplificar la comparación de los espectros como se muestra en la fig. 7B. Tal como se muestra, las características (p. ej., picos) de los dos espectros 702 y 704 están relativamente alineadas, pero tienen diferentes amplitudes. Esta diferencia de amplitud da como resultado las diferentes intensidades de los diferentes espectros 702 y 704. Mediante la normalización de los dos espectros 702 y 704, usando normalización iónica parcial u otro algoritmo de normalización, los dos espectros 702 y 704 se solapan sustancialmente y pueden compararse apropiadamente como se muestra en la gráfica 700b de la fig. 7B.

Las fig. 8A y 8B son las gráficas 800a y 800b que muestran los espectros múltiples de muestra 802a-802n (fig. 8A) que se han alineado 802a'-802n' (fig. 8B). La escala de masa absoluta de los espectros puede variar considerablemente. Los espectros pueden cambiar el uno con respecto a otro, e incluso la escala de masa interna no es constante. En las tareas convencionales de proteómica, se añaden compuestos especiales que dan lugar a picos de valores m/z conocidos. Los espectros pueden recalibrarse después (es decir, los valores m/z pueden reescalarse de acuerdo con estos calibradores externos) y pueden lograrse precisiones de masa absoluta de unas cuantas decenas de ppm en el intervalo inferior de masa, en el que se esperan péptidos. En el caso de las muestras no digeridas, a veces es difícil añadir calibradores al tejido; y suele ser indeseable, ya que los calibradores podrían suprimir picos relevantes debido a efectos de supresión de iones. Para la comparación espectral, sin embargo, es suficiente alinear los espectros con una escala de masas común y no es tan importante que esta escala de masas corresponda realmente con una medida absoluta de la masa (es decir, que no se realizan búsquedas en bases de datos). Puede realizarse la identificación de picos comunes, como se describe con respecto a la fig. 9.

Con el fin de alinear los espectros, pueden identificarse picos comunes a lo largo de los grupos de espectros. Los picos de los espectros se ponen en una línea y pueden usarse algoritmos de agrupamiento divisivo para separar esta larga lista en una lista de grupos de la forma siguiente:

Iniciación: Las posiciones de los picos de los espectros se disponen en una lista ordenada (por valor de m/z)
Primera etapa de separación: En la que puede usarse una separación mínima (típicamente 30 Da) para dividir esta larga lista en grupos de picos, en los que cada pico individual esté más pegado que la separación mínima deseada. Como resultado, puede obtenerse una lista de grupos de picos próximos.
Separación fina: Puede generarse un histograma de diferencias de picos para cada uno de estos grupos. El grupo a la distancia atípica, que se define como el doble de la mediana de la separación de los picos en el grupo puede dividirse, si la distancia dividida es menor que el doble de la anchura del pico o menor que la resolución del equipo en este intervalo de m/z, entonces los grupos no se dividen. Si se produce una división, entonces pude realizarse como recurso el mismo análisis en los dos grupos resultantes hasta que ya no se produzcan más divisiones. Si no se produce ninguna división, entonces se va al siguiente grupo.

Como resultado, se obtiene una lista de grupos que están cercanos por su m/z y bien separados. Cada grupo puede caracterizarse mediante su centro (la mediana de las posiciones m/z de todos los picos del grupo), y su anchura (los percentiles 25º y 75º de estas posiciones). De forma alternativa, pero menos consistente, puede usarse la media y la desviación típica como medida de la localización y diseminación.

Puede realizarse una selección típicamente en el orden de diez grupos de intensidad promedio aceptable diseminados uniformemente a lo largo del intervalo de m/z tanto como sea posible. También puede realizarse una regresión lineal (cuadrática) sobre cada espectro para alinear las escalas de masa de todos los espectros con estos picos comunes. En un ejemplo, pueden usarse los siguientes centros de grupo: 6434,50, 6632,18, 11686,94, 12864,88, 15131,14, 15871,47, 28102,55.
Puede realizarse un alineamiento con una tolerancia de 5000 ppm, es decir, si en algún espectro no se encontró un punto de alineamiento en las posiciones especificadas dentro de esta tolerancia, puede ignorarse este punto. Sin embargo, si un alineamiento no se realiza, los siguientes no se detectan como características: 5764, 8702, 9426, 11443, 11686, 21066, 28102, 28309. Como resultado, la desviación típica mediana de las características se reduce de 4,63 Da a 3,68 Da para los picos que son visibles en los espectros no alineados.

Esta selección de estos picos comunes puede usarse para registrar los espectros en una escala de m/z común. Como se muestra en la fig. 8B.

Extracción de características

Continuando con la fig. 3, se usa un proceso de extracción de características en la etapa 308 para extraer características (p. ej., picos) a partir de los espectros. Al hacerlo, se efectúa una determinación en cuanto a qué características que se van a extraer.

Aunque una inspección visual de los espectros, sus promedios y diferencias de grupos, proporciona cierta orientación sobre la capacidad de distinguir diversos estados o estadios clínicos de la enfermedad usando espectrometría de masas, puede realizarse un análisis más cuantitativo. Un pico de diferenciación se basa en las posiciones m/z de los picos en los espectros. Dicha posición es un marcador provisional si es común a algunos números de espectros definidos por el usuario dentro de un grupo o característica dados. Una vez que se ha creado una lista de estas características para cada grupo, puede darse a cada característica un valor definitorio. Mediante el uso de los ajustes de la anchura del pico de un algoritmo para encontrar picos, las amplitudes normalizadas y con el fondo sustraído pueden integrarse a lo largo de este intervalo y se puede asignar este valor integrado (es decir, el área bajo la curva entre la anchura de la característica) a una característica. Para los espectros en los que no se ha detectado ningún pico dentro de este intervalo m/z, el intervalo de integración puede definirse como el intervalo alrededor de la posición m/z promedio de esta característica con una anchura correspondiente a la anchura del pico en la posición m/z actual.

Los valores de las características pueden variar considerablemente de espectro a espectro, e incluso dentro de la misma muestra (p. ej., suero o tejido), o entre muestras diferentes del mismo tipo celular. Mientras que la posición m/z de los picos es muy reproducible, las amplitudes presentan fluctuaciones considerables.

Como se ha descrito previamente, una determinación de la variación de los valores de características son sus coeficientes de variación (CV). Los coeficientes de variación se definen como el cociente entre la desviación típica de las características y su valor promedio. Son posibles otras definiciones, como el cociente entre el intervalo de los percentiles entre el percentil 25º y el 75º y el valor de su mediana. En un histograma se proporciona una distribución típica de los valores de los CV para los espectros usados. Aunque hay valores de características que son muy reproducibles con valores de CV menores de 0,5, la mayoría de las características muestran una gran variación. Esto pone de relieve por qué la extracción no es trivial y hay que analizar las fluctuaciones y las distribuciones de las características antes de identificar la característica como un pico de diferenciación potencial con una característica distintiva.

Continuando con la fig. 3, se realiza un proceso de selección de característica en la etapa 310 para seleccionar las características que se utilizan en la realización del análisis de clasificación. El proceso de selección de características puede ilustrarse como se muestra en la fig. 9.

La fig. 9 es una gráfica de un proceso de ejemplo para la selección de una característica (características candidatas) mediante la localización de un pico común en más de "x" espectros que tienen una cierta anchura, en el que la anchura se define como un error de alineamiento mas la anchura del pico. Pueden utilizarse diversas técnicas de selección para realizar la selección de característica. Tal como se muestra, hay tres espectros 902a-902c (colectivamente 902). Estos espectros 902a-902c se utilizan para localizar una característica (p. ej., un pico) 904. Tal como se muestra, una línea vertical central 906 se extiende a lo largo del centro de la característica 904, que es común en más de uno de los espectros 902, y las líneas verticales laterales 908a y 908b definen la anchura de la característica (error de alineamiento + anchura del pico).

La selección de las características de diferenciación puede realizarse en un proceso de tres etapas: En primer lugar, todas las características se ordenan mediante un valor de p univariante obtenido a partir de una prueba de hipótesis simple suponiendo que todas las características son independientes. En algunas ejecuciones, puede usarse la prueba de Mann-Whitney para obtener un valor de p para cada característica. Otros métodos son posibles, pero menos consistentes, tales como las pruebas de dos muestras, las pruebas de Kolmogorov Smirnov, u otros. En segundo lugar, mediante el uso de las correcciones de Bonferroni, las características de mayor rango (menor valor de p) se inspeccionan mediante la comparación de los espectros promediados de grupo (el promedio de los espectros de un grupo clínico). Si una característica no distingue grupos, se descarta como candidata. En tercer lugar, y en una etapa final, la selección de característica puede realizarse usando los errores de la validación cruzada como criterio para el éxito. A continuación se destacan diversas ejecuciones a este efecto:

La selección de características relevantes es más que una cuestión de experimentos de micromatriz génica, ya que hay miles de características y pocas muestras. La selección de características es también una cuestión de identificación de biomarcadores cuando se examinan los datos de masa espectral, ya que existen algunas pruebas de que la selección de las características no influye mucho sobre el funcionamiento de algunos clasificadores. No obstante, es difícil interpretar los resultados de la clasificación si hay muchas decenas de características, y, en realidad, no hay expectativas de que todas estas características sean relevantes.

Pueden ordenarse las características por su importancia para diferenciar diversos estadios de enfermedad. Es claro seleccionar una característica cada vez, pero cuando hay muchas decenas de características, la tarea es más difícil para determinar cuáles de las características son las importantes para el estadio particular de la enfermedad. A fin de comparar biomarcadores y espectros a lo largo de los laboratorios, deben ser identificables las mismas características, y aquellas características que parezcan debidas a incertidumbres en la preparación de la muestra, uso del equipo, y variaciones de la población deben ser distinguibles.

La selección de características se enfrenta a dos determinaciones algorítmicas. La primera determinación es puramente combinatoria. Una búsqueda completa de todas las combinaciones posibles de 1 características de un total de m características disponibles (determinadas) conduce a $\binom{m}{l} = \frac{m!}{l!(m-l)!}$ combinaciones, p. ej., para m= 20, 1=5 este número es 15504. Como, típicamente, en los espectros de masas, hay un par de cientos de características disponibles, este número de combinaciones puede ser demasiado grande para una búsqueda completa. Asimismo, puede no ser fácilmente aparente qué valor de 1 sea óptimo. Por tanto, pueden usarse estrategias especiales de búsqueda heurística. El segundo mecanismo de determinación surge a partir de la falta de una determinación de calidad exclusiva que decida qué conjunto de características es mejor que otro. Como un criterio para la selección de características podría ser la realización de la clasificación, los "métodos envolventes" incorporan la selección de características como parte del algoritmo de clasificación. Estos métodos usan una estimación del error de clasificación, en el mejor de los casos, la determinación del error de generalización, que es difícil de determinar, y al que se aproxima típicamente mediante validación cruzada dejando uno fuera (LOOCV, de *leave-one out cross-validation*), o límites de error basados en el margen en el caso del aprendizaje de máquinas de soporte vectorial (MSV). Las alternativas incluyen métodos de filtro que realizan la selección de características antes de que se generen los clasificadores. Cada uno de estos planteamientos tiene sus propias cuestiones, y utiliza un manejo especial con respecto a la validación.

Las estrategias de búsqueda se discuten más adelante en primer lugar, y después se enumera un conjunto de determinaciones de calidad que se usan habitualmente.

Estrategias de búsqueda de características

La mayoría de las estrategias de búsqueda se basan en un planteamiento de "divide y vencerás", que optimiza el criterio de selección de característica. Para las elecciones específicas del criterio de selección de características, puede ser posible usar muestreo probabilístico en la esencia de la importancia del muestreo de Monte Carlo, o de técnicas especiales de optimización, tales como programación dinámica.

Tal como se usa, el agrupamiento en árbol puede empezar con todas las características y las características pueden borrarse una por una. De forma alternativa, el proceso puede empezar con una característica y añadir otras características una por una. Como ilustración, pueden existir cuatro características $\{x_1,x_2,x_3,x_4\}$.

Búsqueda de arriba a abajo:

- Calcular el valor del criterio de selección de características para $\{x_1,x_2,x_3,x_4\}$ que da C_4.
- Calcular el valor del criterio de selección de características para cada una de $\{x_1,x_2,x_3\}$, $\{x_1,x_2,x_4\}$, $\{x_1,x_3,x_4\}$, $\{x_1,x_2,x_4\}$, y seleccionar el mejor, digamos $\{x_1,x_2,x_3\}$ con el valor C_3.
- Calcular el valor del criterio de selección de características para cada una de $\{x_1,x_2\}$, $\{x_1,x_3\}$, $\{x_2,x_3\}$, seleccionar el mejor, digamos $\{x_1,x_2\}$ con el valor C_2.
- Y, finalmente, recoger la mejor característica individual de entre $\{x_1,x_2\}$ con el valor C_1.
- El mejor valor de $\{C_1,C_2,C_3,C_4\}$ define el conjunto de características (sub)óptimo.

Empezando de forma similar a partir de una característica, y añadiendo más, una por una, se define una búsqueda de abajo a arriba. Esto no da necesariamente una solución óptima, ya que no hay una garantía de que el número óptimo menor (mayor) de solución de características evolucione de acuerdo con estos árboles. Una forma de mejorar estos sencillos procedimientos es reconsiderar características descartadas previamente, o descartar características seleccionadas previamente. Este algoritmo se llama método flotante de búsqueda, tal como se entiende en la técnica, y de la forma siguiente:

Método flotante de búsqueda:

A continuación se describe una búsqueda para un número fijo 1 de m características. Puede realizarse un bucle sobre 1 para optimizar el número de características. El método flotante de búsqueda se basa tanto en búsquedas de arriba a abajo como de abajo a arriba. El algoritmo descrito se basa en el método de abajo a arriba.

Considérese un conjunto de m características. La idea es buscar el mejor subconjunto de k de ellas para $k = 1,2,...,l \leq m$ optimizando C. Sea $X_k = \{x_1,...,x_k\}$ el conjunto óptimo para K características, e Y_{m-k} el conjunto de las restantes m-k características. Los mejores subconjuntos dimensionales $X_2,X_3,...,X_{k-1}$, de 2,3,...,k-1 características se mantienen almacenados. En la siguiente etapa, el (k+1)º subconjunto óptimo X_{k+1} se forma tomando un elemento de Y_{m-k}. Después, se realiza una comprobación a lo largo de todos los subconjuntos dimensionales en cuanto a si esto mejora C, y sustituye a la característica seleccionada previamente. El algoritmo funciona de la forma siguiente (C mejor cuanto más grande).

- Seleccionar la mejor característica individual, que dé X_1 con C_1.

- Añadir otra basada en C, que dé X_2 y C_2.

Ahora, iterar sobre k:

- Etapa I, Inclusión: elegir aquel elemento de Y_m-$_k$ que combinado con X_k dé el mejor C, es decir, X_{k+1} = arg $\max_{y \in Ymk} C(X_k, y)$ definiendo $X_{k+1} = \{X_k, x_{k+1}\}$ como en el algoritmo de abajo a arriba.
- Etapa II, Análisis:

  1. Encontrar la característica x_r que tenga el menor efecto sobre el coste C cuando se elimine de X_{k+1}, *es decir,* x, = arg $\max_{xr \in k+1} C(X_{k+1}/\{x_r\})$.
  2. Si $r = k + 1$, $k = k + 1$, $C_{k+1} = C$ e ir a la etapa I.
  3. Si $r \neq k + 1$ y $C(X_{k+1}/(x_r)) < C_k$ ir a la etapa I, es decir, si la eliminación de x_r no mejora el grupo seleccionado previamente, no hacer una búsqueda hacia atrás.
  4. Caso especial para k=2: Si k=2 ajustar $X_2 = X_3/\{x_r\}$ y $C_2 = C(X_3/\{x_r\})$.

- Etapa III, Exclusión (búsqueda hacia atrás):

  1. $X_k' = X_{k+1}/\{x_r\}$ es decir, eliminar x_r
  2. Encontrar la característica menos significativa x_s' en el nuevo conjunto por medio de $x_s = \arg\max_{y \in X_k'} C(X_k'/\{y\})$.
  3. Si $C(X_k'/\{x_s\}) < C_{k-1}$ entonces $X_k = X_k'$, reajustar C_k e ir a la etapa I terminando la búsqueda hacia atrás.
  4. Ajustar $X_{k-1}' = X_k'/\{x_s\}$ y k=k-1.
  5. Caso especial k=2: Ajustar $X_2 = X_2'$ y $C_2 = C(X_2')$ e ir a la etapa I.
  6. Ir a la etapa III.

Este algoritmo funciona, por lo general, sustancialmente mejor que el simple algoritmo de abajo a arriba, y puede efectuarse hasta m para recoger de nuevo el máximo (mínimo) conjunto de criterios.

Algoritmo de selección aleatoria de características

El algoritmo de selección aleatoria de características es una estrategia de optimización basada en el recuento de la frecuencia de configuraciones a partir de un muestreo al azar. Por ejemplo, al construir grupos jerárquicos aglomerativos a partir de alguna configuración iniciar (medianas de k, medias de k, agrupamiento difuso), el algoritmo puede iniciarse muchas veces más, almacenar las configuraciones individuales de cada sesión, y construir un histograma de frecuencias. A menudo, esto puede combinarse con validación cruzada.

Generación de clasificador

Continuando con la fig. 3, en la etapa 312, se realiza una generación de clasificador. La generación de clasificador puede incluir unas cuantas funciones, que incluyen (i) aprendizaje supervisado, (ii) validación cruzada, y (iii) clasificación o análisis con ocultación. Las primeras dos funciones, aprendizaje supervisado y validación cruzada, pueden realizarse sobre los espectros brutos con los resultados clínicos asociados conocidos 108 proporcionados por las clínicas de investigación del cáncer 104, como se describe en la fig. 1.

Aunque el orden de las características da alguna idea sobre la importancia de las características para la discriminación del os grupos, se usa un análisis más exhaustivo en un procedimiento de aprendizaje supervisado. El aprendizaje supervisado es el proceso mediante el cual se proporcionan etiquetas de categoría para cada caso, en un conjunto de prácticas (es decir, cada espectro) y busca reducir el número de clasificaciones erróneas. Otra definición más específica de aprendizaje supervisado es la cartografía a partir de un espacio de característica de alta dimensionalidad para etiquetar el espacio desde la expresión de la característica/pico de diferenciación hasta la etiqueta de enfermedad o la etiqueta de respuesta (designada de otro modo como etiqueta de clase). La etiqueta es una función de los picos del espectrómetro de masas y los parámetros asociados. Un investigador u otra persona que tenga espectros de, e información clínica sobre, el paciente con cáncer a partir del cual se produjo el espectro, puede realizar el proceso de aprendizaje supervisado. El proceso puede realizarse usando algoritmos convencionales de la teoría del aprendizaje supervisado. El resultado de los algoritmos de clasificación supervisada es un algoritmo clasificador (dependiente del conjunto de prácticas) que genera una etiqueta de clase para un nuevo caso o espectro. En un ejemplo, puede utilizarse un algoritmo de los vecinos más próximos de k (VMPK) para la clasificación.

Algoritmo de los vecinos más próximos de k

El método de los vecinos más próximos de k es un método sencillo de estimación de la densidad. La probabilidad de que un punto x' esté clasificado dentro de un volumen V centrado en x es:

$$p \approx \int_V dx\, p(x)$$

Para un volumen pequeño $p \approx p(x)V$. La probabilidad puede aproximarse mediante la proporción de muestras que se clasifican dentro del volumen V. Por lo tanto, si k es el número de muestras fuera de un total de n que se clasifican dentro de V, entonces

$$p \approx \frac{k}{n} \quad \text{y} \quad p(x) = \frac{k}{nV}$$

La aproximación del vecino más próximo de k es para fijar la probabilidad k/n (o para un número fijo de muestras para fijar k) y determinar el volumen que contiene k muestras. Esto está en contraposición con las estimaciones de histograma, en las que la anchura del intervalo es fija, y el número de puntos se cuenta. Hay algunos problemas con la regularidad de esta definición, pero puede demostrarse que no está sesgada y es coherente si $\lim_{n\to\infty} k(n) = \infty$ y $\lim_{n\to\infty} k(n)/n = 0$.

Puede construirse una regla de decisión de la siguiente forma. Supóngase que hay k_m muestras en la clase ω_m, y que el número total de muestras de ω_m es n_m. Entonces, la probabilidad condicionada de clase es:

$$p(x \mid \omega_m) = \frac{k_m}{n_m V}$$

La previa es n_m/n (si hay n muestras en total a lo largo de todas las clases).

La regla de decisión bayesiana es asignar x a ω_m si

$$p(\omega_m \mid x) \geq p(\omega_i \mid x)\ \forall i$$

y usar el teorema de Bayes; esto da como resultado esta selección

$$\frac{k_m}{n_m V}\frac{n_m}{n} \geq \frac{k_i}{n_i V}\frac{n_i}{n}\ \forall i \Rightarrow k_m \geq k_i$$

En caso de empate, puede efectuarse un desempate por medio de la media más próxima, el miembro más cercano, o de otra forma. De forma alternativa, el desempate puede restringirse al k impar. Un k pequeño conduce a superficies irregulares, mientras que un k grande, a superficies lisas. El índice asintótico de clasificación errónea se acota desde arriba mediante el doble del error de Bayes, que es un rendimiento asintótico muy bueno para un algoritmo tan sencillo. La clasificación VMPK conduce por sí misma al uso de prototipos, es decir, una técnica de condensación de datos. Pero en el presente documento, el uso de la clasificación VMPK se usa más para la reducción en el almacenamiento necesario. Puede utilizarse la elección de una función de distancia. De forma alternativa, también pueden utilizarse las diferencias euclidianas, lo cual no es óptimo. El proceso de votación para un ejemplo sencillo de un espacio bidimensional de características se ilustra en la fig. 11.

La fig. 11 es una gráfica 1100 que muestra un grupo de indicios de espectros etiquetados por clase de ejemplo representativos de dos clases diferentes de indicios de avance de la enfermedad y de espectro de análisis para clasificar. Para representar gráficamente los picos de diferenciación en el espacio de características, en esta ilustración de un espacio bidimensional de características, la gráfica 1100 es una gráfica bidimensional que tiene un eje x y un eje y. Si el espacio de características fuese un espacio de características de 12 dimensiones (es decir, que se seleccionasen 12 características o picos como picos de diferenciación indicativos o que distinguen características que clasificasen un espectro para ser etiquetado por clase como "bueno" o "malo") entonces no sería posible representar gráficamente los espectros con facilidad, por lo que se utiliza un espacio de características bidimensional como ejemplo.

En este caso, los espectros se clasifican con las etiquetas de clase como "bueno" 1102 y "malo" 1104, en las que los indicios de los espectros con la etiqueta de clase "bueno" 1102 se representan en la gráfica 1100 como un patrón y los indicios de los espectros con la etiqueta de clase "malo" 1104 se representan como otro patrón. Como se ha descrito previamente, pueden desarrollarse espectros con etiquetas de clase de una clínica de investigación del cáncer y usarse como muestra de control con fines de clasificación basada en los resultados clínicos de un paciente con cáncer al responder a un fármaco anticanceroso, tal como Iressa. Pueden colocarse los indicios 1106 de un espectro de análisis en la gráfica 1100 en una localización representativa de un espectro de análisis de un nuevo paciente con cáncer para el que se está determinando un plan de tratamiento. La localización de los indicios 1106

del espectro de análisis se basa en las amplitudes de las dos características (es decir, las amplitudes x e y). Tal como se muestra, y de acuerdo con el algoritmo de probabilidad VMPK, los tres indicios de espectros etiquetados por clase más cercanos 1108a, 1108b, y 1108c son candidatos potenciales para el espectro de análisis que se va a asociar.

Un análisis de probabilidad de ejemplo para el proceso de clasificación para un punto de análisis del espacio bidimensional de característica es:

$$P(\vec{x} \in A) = \frac{k_A + 1}{k_A + k_B + 2} \; {}_2F_1(1, k_B + 1, k_A + k_B + 3, 1 - \frac{N_A}{N_B})$$

Si la diferencia de probabilidad entre dos clases excede un cierto umbral delta-p suministrado por el usuario, entonces la probabilidad puede considerarse significativa y puede efectuarse una clasificación de "bueno" o "malo". Si la diferencia de probabilidad está por debajo de un cierto umbral, entonces puede efectuarse una clasificación de "indeterminado".

Aunque puede utilizarse un algoritmo VMPK como algoritmo clasificador, pueden utilizarse otros algoritmos de clasificación. Otro algoritmo desarrollado de acuerdo con los principios de la presente invención es un algoritmo probabilístico del vecino más próximo de k, que es un algoritmo VMPK modificado que proporciona flexibilidad adicional y proporciona más información para las aplicaciones clínicas.

Algoritmo modificado (probabilístico) del vecino más próximo de k

De acuerdo con los principios de la presente invención, puede usarse un algoritmo modificado de los vecinos más próximos de k para la clasificación. En su ejecución más simple, el algoritmo VMPK modificado busca los vecinos más próximos de k en el espacio de características y asigna una etiqueta de clase de acuerdo con un voto de mayoría simple sobre las etiquetas de estos vecinos más próximos El espacio de características se define como que es el número de características (p. ej., 12 características) que se están usando para definir un espectro. En un ejemplo, no hay una fase de prácticas explícita y se usan todos los casos en la clasificación de los espectros. Normalmente, sólo se usan las distancias euclidianas simples para determinar los vecinos, pero son posibles otras definiciones (p. ej., las distancias de Mahanolobis a partir de matrices de covarianza definidas de forma apropiada).

En el marco tradicional de los vecinos más próximos de k (VMPK), la clasificación se realiza de la forma siguiente:

Cada objeto, o caso, que se va a clasificar (aquí - el espectro de masa) se caracteriza mediante los números $d\ x_i$, I = 1 ....D (aquí- los valores de d características), y se representan, por tanto, mediante un punto en el espacio d-dimensional. La distancia

entre los dos casos se define mediante la métrica euclidiana habitual $\sqrt{\sum_i (x_i - x_i')^2}$. Por supuesto, aquí puede usarse también cualquier métrica similar. De forma adicional, una ejecución puede usar una distancia de Mahalanobis winsorizada al determinar la distancia entre dos espectros.

Un conjunto de prácticas puede incluir casos con asignaciones de clase conocidas. Dado el conjunto de prácticas y un entero positivo impar k, la clasificación del objeto de análisis se realiza de la forma siguiente:

1. En el conjunto de prácticas, encontrar los vecinos más próximos de k del objeto de análisis (es decir, el espectro) en el espacio d-dimensional.
2. Cada uno de estos vecinos de k pertenece a una de las clases (p. ej., bueno o malo). Encontrar qué clase tiene el número mayor de representantes.
3. Clasificar el objeto de análisis como perteneciente a esta clase.

Esta clasificación VMPK tiene dos inconvenientes. En primer lugar, no proporciona ninguna información sobre la confianza de la asignación de clase. De forma intuitiva, está claro que en el caso de k= 15 y dos clases, la confianza de la asignación de clase en la situación 15: 0 es mucho mayor que en la situación 8: 7. En las aplicaciones clínicas, se caracteriza porque el nivel de confianza de cada asignación de clase individual es relevante y se usa para diagnosticar a los pacientes. De hecho, este nivel puede definirse al principio.

En segundo lugar, no tiene en cuenta de manera apropiada el número de casos de cada clase en el conjunto de prácticas. La sola adición de más casos de la clase dada al conjunto de prácticas tiende a sesgar los resultados de la clasificación en favor de esta clase.

Para corregir estos problemas, se ha desarrollado un clasificador "VMPK probabilístico" que parte de la información de las clases de los vecinos más próximos de k del conjunto de prácticas, pero, en lugar de asignación de clase produce probabilidades de que el caso de análisis pertenezca a cada una de las clases. Más adelante hay una descripción concisa del razonamiento y la procedencia de las fórmulas principales para el VMPK probabilístico.

El planteamiento VMPK para la clasificación de las muestras de espectro puede verse de la forma siguiente: Considérese una pelota de un cierto radio en el espacio d-dimensional y centrada en el caso de análisis. El radio de la pelota está determinado por el requisito de que contiene exactamente k casos del conjunto de prácticas. Después, obsérvese cuántos miembros de cada clase están entre estos k casos, y úsese esta información para asignar la etiqueta de clase (en el planteamiento convencional), o compútense las probabilidades de que el caso de análisis pertenezca a esta o aquella clase (en el planteamiento probabilístico).

El conjunto de prácticas puede ser una muestra extraída de una cierta (desconocida) distribución de probabilidad. De forma más precisa, para cada clase, se considera que el subconjunto del conjunto de prácticas que pertenece a la clase es una muestra extraída de la correspondiente distribución de probabilidad, la cual es diferente para cada clase.

Considérese el grupo de conjuntos de prácticas extraídos de la misma distribución de probabilidad. En el planteamiento VMPK para la clasificación, el radio de la pelota alrededor del caso de análisis es diferente para cada realización del conjunto de prácticas para asegurar que siempre contenga exactamente k vecinos más próximos. Véase también la descripción del método VMPK de la sección anterior.

Pueden efectuarse las siguientes aproximaciones:

1. La pelota alrededor del caso de análisis puede considerarse fija, lo que significa, que es dependiente de la posición del caso de análisis y de las distribuciones de probabilidad a partir de la cual se ha extraído el conjunto de prácticas, pero es la misma para cada realización del conjunto de prácticas. Esta aproximación es válida cuando k no es demasiado pequeño.
2. Para cada clase, el número de casos de esta clase dentro de la pelota se extrae de la distribución de Poisson. Esta aproximación es válida cuando la pelota contiene sólo una pequeña fracción de la probabilidad global para esta clase.
3. Las densidades de probabilidad para las clases son aproximadamente constantes dentro de la pelota.

Considérese el caso de dos clases. Cada caso se representa mediante un punto x en el espacio *d*-dimensional. El espacio *d*-dimensional completo se señala mediante Ω.

La clase 1 se caracteriza por la distribución de probabilidad $p_1(\bar{x})$, $\int_{\Omega} p_1(\bar{x})d\bar{x} = 1$. La clase 2 se caracteriza por la distribución de probabilidad $p_2(\bar{x})$, $\int_{\Omega} p_2(\bar{x})d\bar{x} = 1$.

Un conjunto de prácticas puede estar formado por N_1 puntos extraídos de la clase 1, y N_2 puntos extraídos de la clase 2. La proximidad del punto de análisis puede señalarse mediante ω. Ésta es realmente una pelota centrada en el punto de análisis, pero es irrelevante para lo siguiente Para una realización dada del conjunto de prácticas, hay *k1* puntos en ω desde la clase 1 y k_2 puntos en in ω desde la clase 2. Se supone que $k_1 << N_1$. $\int_{\omega} p_1(\bar{x})d\bar{x} << 1$ Lo mismo es cierto para la clase 2.

Esto asegura la validez de la aproximación de Poisson: k_1 viene de la distribución de Poisson con el valor esperado λ_1,

$$\lambda_1 = N_1 \int_{\omega} p_1(\bar{x})d\bar{x};$$

k_2 viene de la distribución de Poisson con el valor esperado λ_2,

$$\lambda_2 = N_2 \int_{\omega} p_2(\bar{x})d\bar{x}.$$

Ahora el punto de análisis (el centro de *ω*) se trata como "otro punto más". En otras palabras, hay $k_1 + k_2 + 1$ puntos en *ω*, en vez de k_1+k_2, y no se sabe a qué clase pertenece el punto de análisis. Las probabilidades del punto de análisis de pertenecer a la clase 1 y a la clase 2 pueden asignarse de la forma siguiente:

$$\frac{p(class1)}{p(class2)} = \frac{\int_{\omega} p_1(\bar{x})d\bar{x}}{\int_{\omega} p_2(\bar{x})d\bar{x}}$$

Por tanto

$$p(class1) = \frac{\int_{\omega} p_1(\bar{x})d\bar{x}}{\int_{\omega} p_1(\bar{x})d\bar{x} + \int_{\omega} p_2(\bar{x})d\bar{x}} = \frac{\frac{\lambda_1}{N_1}}{\frac{\lambda_1}{N_1} + \frac{\lambda_2}{N_2}}$$

Tratando el punto de análisis (el centro de ω) como "otro punto más" se supone de forma implícita que ni $p_1(\bar{x})$ *ni* $p_2(\bar{x})$ cambian de forma significativa dentro de ω.

El problema es que λ_1 y λ_2 son realmente desconocidos. Sus probabilidades, sin embargo, pueden estimarse de la manera bayesiana. Se supone que tanto k_1 como k_2 obedecen a la distribución de Poisson,

$$p(k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Señalando la distribución anterior para λ mediante $p_0(\lambda)$,

$$p(k) = \int d\lambda p(k \mid \lambda) p_0(\lambda).$$

Mediante el razonamiento bayesiano convencional,

$$p(\lambda \mid k) = \frac{p(k \mid \lambda) p_0(\lambda)}{\int d\lambda p(k \mid \lambda) p_0(\lambda)}.$$

Suponiendo a partir de ahora que la distribución categórica previa de λ, $p_0(\lambda) = 1$, puede obtenerse lo siguiente

$$p(\lambda \mid k) = p(k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Finalmente, se obtiene lo siguiente

$$p(class1) = \int_0^{\infty} d\lambda_1 \int_0^{\infty} d\lambda_2 \frac{\lambda_1}{\lambda_1 + \frac{N_1}{N_2}\lambda_2} p(\lambda_1) p(\lambda_2),$$

donde

$$p(\lambda_1) = \frac{\lambda_1^{k_1}}{k_1!} e^{-\lambda_1},$$

$$p(\lambda_2) = \frac{\lambda_2^{k_2}}{k_2!} e^{-\lambda_2}.$$

La computación de estas integrales da lo siguiente:

$$p(class1) = \frac{k_1 + 1}{k_1 + k_2 + 2} \, {}_2F_1(1, k_2 + 1, k_1 + k_2 + 3, 1 - \frac{N_1}{N_2}).$$

Para tamaños iguales de las muestras usadas en el conjunto de prácticas (N_1=N_2), esto simplifica lo siguiente:

$$p(class1) = \frac{k_1 + 1}{k_1 + k_2 + 2},$$

$$\frac{p(class1)}{p(class2)} = \frac{k_1 + 1}{k_2 + 1}.$$

Para más de dos clases y diferentes tamaños en el conjunto de prácticas, es difícil obtener p(clase I) en forma cerrada. En este caso, puede usarse la siguiente estimación, muy simplificada:

$$\frac{p(class\ i)}{p(class\ j)} = \frac{k_i + 1}{k_j + 1} \cdot \frac{N_j}{N_i},$$

O, de forma equivalente, cada p(clase l) es proporcional a $\frac{1}{N_i}(k_i+1)$, mientras que $\sum_{i=1}^{N_{classes}} p(class \quad i) = 1$.

El parámetro que caracteriza la consistencia de los resultados para los espectros defectuosos es un parámetro suministrado por el usuario, p-dif, que define en qué medida han de ser diferentes las probabilidades de clase a fin de asociar una etiqueta con un espectro. Por ejemplo, si p-dif se ajusta a 0,1 y la probabilidad para la clase A es 0,6 y la probabilidad para la clase B es 0,4, entonces la diferencia 2 es mayor de 0,1, y se elegirá la clase A. Si, por otra parte, la probabilidad para la clase A es 0,52 y la probabilidad para la clase B es 0,48, entonces la diferencia 0,04 es menor de 0,1, y el clasificador devuelve una etiqueta de clase que es "No definido".

De forma alternativa, el análisis de la hipótesis puede dar una clasificación significativa con una significación α especificada de forma externa. En una formulación convencional del análisis de la hipótesis, la clasificación puede describirse de la forma siguiente:

Datos: Un caso de análisis puede incluir dos clases, A y B, los vecinos más próximos de k_A y k_B de la clase A y la clase B, y las poblaciones de N_A casos de la clase A y N_A casos de la clase B.

Estadística del análisis: Solamente el número de vecinos de la clase A:

$$T = k_A$$

Distribución nula: Se supone que el nulo es el número de vecinos de A que se espera a partir de las proporciones de la población solamente, es decir, que k_A por debajo del nulo es una variable aleatoria binomial con los parámetros $k = k_A + k_B$ y $p^* = N_A/N_B$.

Hipótesis: (bilateral). Ésta es una ejecución de un análisis binomial, tal como se entiende en la técnica.

$$H_0 : p_A = p^*$$

$$H_1 : p_A \neq p^*$$

En el caso del desarrollo del análisis, el número de vecinos más próximos raramente es mayor de 20, por lo que el uso de la aproximación binomial no se emplea. Para una significación global dada α se resuelve a partir de una tabla (o se utiliza un ordenador) $P(Y \leq t_1) = \alpha_1$ y $P(Y \leq t_2) = 1 - \alpha_2$ para t_1 y t_2, donde Y es una variable aleatoria binomial, como se define bajo el nulo, y donde α_1 y α_2 son aproximadamente a $\alpha/2$ y se añaden hasta α. Las regiones de rechazo son los valores de T menores de t_1 o mayores de t_2. También pueden estimarse las regiones de confianza para p^* siguiendo el procedimiento destacado en la sección Análisis binomial.

Aunque el algoritmo VMPK puede utilizarse como el algoritmo de clasificación como se ha descrito anteriormente, pueden utilizarse algoritmos de clasificación alternativos de acuerdo con los principios de la presente invención. Dichos algoritmos de clasificación pueden incluir VMPK difuso, métodos de Kernel (p. ej., MVS), clasificación no supervisada, agrupamiento espectral, ACP de Kernel, agrupamiento no paramétrico, medias de k, histogramas de k agrupamiento jerárquico, y bosques aleatorios, por ejemplo. Estos algoritmos de clasificación proporcionan la capacidad de clasificar un espectro de acuerdo con espectros etiquetados por clase (p. ej., espectros que se han clasificado y etiquetado a partir de un grupo de control de pacientes con cáncer), pero carecen de la transparencia y facilidad de uso de los algoritmos VMPK descritos anteriormente.

Continuando con la fig. 3, etapa 312, puede utilizarse el aprendizaje para la generación de clasificadores para un conjunto espectros de prácticas. En caso de muestreo de suero para detectar si un fármaco anticanceroso será eficaz sobre el carcinoma no microcítico de pulmón, se utilizaron grupos de control de pacientes, que incluían el uso de tres conjuntos de pacientes cuyos cánceres avanzaron después de la quimioterapia. Se trató a cada uno de los pacientes con Iressa y se registró la información, incluyendo los tiempos de supervivencia de estos pacientes. Las muestras de control fueron de pacientes con casos menos graves (estadios III y IV de cáncer) que no recibieron tratamiento con inhibidores de EGFR-K1, y se produjo suero durante el tratamiento. En la Tabla III se proporciona un resumen de los conjuntos de datos usados en varios estudios. Cada conjunto de datos representa al centro de investigación del cáncer del que se recibieron los espectros y la información asociada de los pacientes.

TABLA III Conjuntos de datos usados en el estudio

Conjunto de datos	Tamaño	Datos del paciente	Uso
Italiano 1	70	Completos	Conjunto de prácticas para la fig. 12, Conjunto de análisis para la fig. 13

Japón 1	43	Pronóstico Supervivencia,	Conjunto de prácticas para la fig. 13, Conjunto de análisis para la fig. 12
Japón 2	26	Pronóstico Supervivencia,	Conjunto de prácticas para la fig. 13, Conjunto de análisis para la fig. 12
VUMC	100	Supervivencia,	Conjunto de control
Italiano 2	69	Supervivencia,	Conjunto de análisis con ocultación total para la fig. 14

La Tabla III es un resumen de los atributos de los conjuntos de datos usados en un estudio para determinar si un algoritmo clasificador podría ser eficaz para determinar si un paciente con cáncer sería de respuesta positiva a Iressa. Los conjuntos de datos italiano 1, italiano 2, Japón 1, y Japón 2 se trataron con Iressa después de la recogida de la muestra. Las prácticas y el análisis en la fase de desarrollo se hicieron de forma transversal a partir de los conjuntos italiano 1 y los dos conjuntos japoneses. Los datos de pacientes incluyeron datos de supervivencia, en los que el conjunto italiano tuvo una historia clínica muy completa junto con el tratamiento y el tipo de cáncer, los conjuntos japoneses sólo incluyeron información pronóstica relacionada con las definiciones de la OMS de etiquetas clínicas, que incluyen enfermedad estable (EE), enfermedad progresiva (EP), y de respuesta parcial (RP) determinada mediante TC. Una vez establecido el clasificador, se realizó un análisis con ocultación total sobre el conjunto italiano 2.

La fig. 10A es una gráfica 1000a representativa de un proceso de ejemplo para clasificar un espectro de análisis en relación con un grupo de espectros etiquetados por clase, de acuerdo con los principios de la presente invención. Se considera que un espectro de análisis tiene una relación con unos espectros etiquetados por clase si se determina mediante un clasificador que el espectro de análisis va a etiquetarse por clase igual que al menos un espectro etiquetado por clase de entre los espectros etiquetados por clase. Las curvas son espectros promediados por grupo. Tal como se muestra, hay un grupo de picos de diferenciación alrededor de 11700 Daltons (Da) usado en la clasificación. Las diferencias entre los grupos están entre los promedios de los espectros de los grupos con etiqueta clínica fina EP precoz 1002 y EE larga 1004. Aunque no se muestran, hay 11 picos de diferenciación usados para construir un clasificador (es decir, un algoritmo clasificador usando el clasificador modificado del vecino más próximo de k) a partir de los conjuntos de datos italianos (tabla III), y sus parámetros se optimizan usando validación cruzada. Está claro al comparar los dos espectros promediados por grupos que una presencia de los biomarcadores resultantes en los picos de diferenciación de los espectros de los pacientes con cáncer de avance rápido (EP precoz 1002) está casi ausente en los pacientes que sobreviven durante largo tiempo y se han clasificado con cáncer de EE larga (EE larga 1004).

Las figs. 10B y 10C son las gráficas 1000b y 1000c que muestran gráficos de ejemplo de los conjuntos de prácticas italiano y los dos japoneses. En la fig. 10B, la gráfica 1000b oscila desde 5500-6000 Da y en la fig. 10c, la gráfica 1000c oscila desde 11000-13000 Da. Tal como se muestra en estas dos gráficas 1000a y 1000b, se muestran numerosos picos de diferenciación entre los diferentes grupos. Los gráficos de los grupos están promediados a lo largo de cada grupo de espectros Esto es, los gráficos no son espectros individuales.

La inusual clasificación fina de los patrones de los picos de diferenciación se refleja de hecho en la fuerza de los picos de diferenciación indicados. En la tabla IV se muestra una lista de los picos de diferenciación usados. La tabla V es la misma lista de picos de diferenciación de la tabla IV, pero también incluye valores de características que contienen promedios de grupo de los valores de características para la fase de descubrimiento de las muestras (italiano 1, japoneses 1 y 2). En la fig. 10 se muestra un conjunto de los grupos dominantes en forma de promedios de grupo. Debe entenderse que los picos de diferenciación que se muestran son de ejemplo y pueden utilizarse los mismos u otros picos de diferenciación de acuerdo con los principios de la presente invención para predecir los pacientes con cáncer que responderán al fármaco Iressa. Más aún, si fueran a efectuarse predicciones para otros anticancerosos u otros fármacos, pueden utilizarse otros picos de diferenciación aparte de los enumerados para dichas predicciones.

El clasificador VMPK óptimo da como resultado un error de validación cruzada dejando uno fuera (LOOCV), mientras que los espectros 6 a 26 no pudieron clasificarse. Mediante el incremento de los requisitos para los clasificadores VMPK probabilísticos, es posible mover este etiquetado erróneo al caso de uno de los espectros inclasificables. Si se supone razonablemente que la clasificación fina se correlaciona con el pronóstico, siendo los casos de EP precoz los de peor avance y los casos de EE larga los de las enfermedades estables más largas, puede concluirse de forma provisional que es posible obtener información prospectiva de respuesta farmacológica a partir de los espectros de suero previos al tratamiento.

TABLA IV Lista de picos de diferenciación

mz_central	mz_bajo	mz_alto	anchura= mz alto-mz bajo

5763,791	5732,131	5795,45	63,3
5843,241	5811,097	5875,384	64,3
6433,973	6398,186	6469,759	71,6
11445,75	11376,15	11515,34	139,2
11529,52	11459,32	11599,73	140,4
11685,37	11614,03	11756,71	142,7
11759,16	11687,28	11831,04	143,8
11903,24	11830,3	11976,18	146,9
12452,38	12375,37	12529,4	154
23354,35	23183,57	23525,13	341,6
23451,25	23279,53	23622,97	343,4
66702,45	65902,02	67502,88	1600

TABLA V. Lista de los picos de diferenciación que contienen parámetros de valores de características

m/z	Valor (Bueno)	Patrón (Bueno)	Valor (Malo)	Patrón (Malo)	Anchura
5763,791	25,387	11,038	113,79	129,02	63,3
5843,241	22,617	11,595	120,27	199,13	64,3
6433,973	402,09	142,69	397,01	165,53	71,6
11445,75	22,334	16,645	353,57	756,68	139,2
11529,52	36,524	39,911	951,3	1401,1	140,4
11685,37	40,505	43,465	1019,9	2136,8	142,7
11759,16	29,745	22,773	341	472,6	143,8
11903,24	20,727	9,2393	158,1	290,08	145,9
12452,38	16,825	10,226	73,804	83,106	154
23354,35	31,089	12,447	63,381	39,39	341,6
23451,25	28,718	13,185	55,475	31,4	343,4
66702,45	342,98	250,02	369,86	203,21	1600

Al analizar el algoritmo clasificador, pueden elaborarse marcadores de respuesta para Iressa con las siguientes asociaciones: Los casos de EE y RP se agrupan juntos en un grupo que tiene una etiqueta de clase de "bueno" y los casos de EP se etiquetan por clase como "malo". El clasificador desarrollado a partir de la clasificación fina anterior fue entonces, de nuevo, la asociación de "bueno" con EE larga y "malo" con EP precoz. Este clasificador se aplicó después a los casos japoneses (tabla I), en los que 18 de estos espectros no pudieron clasificarse, dejando 51 espectros para su clasificación. De estos 51 espectros, 37 tuvieron la etiqueta de clase "bueno" y 14 tuvieron la etiqueta de clase "malo". Los resultados del análisis se resumen en la tabla VI

TABLA VI. Etiquetas de clase

Resultado del análisis	Etiqueta original de clase "bueno"	Etiqueta original de clase "malo"
"bueno"	32	6
"malo"	5	8

Este análisis tiene una sensibilidad del 90 % y una especificidad del 57 %. Para los fines del uso de Iressa, 6 casos, en los que no hubo ninguna respuesta, es decir "malos" se etiquetaron como de respuesta, y dieron un valor

predictivo positivo de 0,84 De forma similar, 5 casos se clasificaron erróneamente como "malos" y dieron un valor predictivo negativo de 0,61.

Resumiendo, mediante el uso de un análisis con espectrómetro de masas basado en suero para filtrar a los pacientes que no responden de los que responden en el grupo de población japonesa se incrementa el índice de respuesta a Iressa de 65 % hasta 90 %, aunque se hubiera dejado fuera a 5 de 51 pacientes, que podrían haberse beneficiado de Iressa. De estos 5 pacientes, uno se etiquetó como EE y 4 se etiquetaron como RP. En general, la clasificación en EP es la peor debido a la alta variabilidad en este grupo. Esto no influye en la selección de los casos "buenos", pero da como resultado la baja especificidad. Este incremento indica que un facultativo podría obtener predicciones inesperadamente mejores del pronóstico del uso de Iressa de forma precoz en la etapa de tratamiento para un cierto grupo de pacientes. Para estos pacientes, se podría continuar con Iressa, mientras que los pacientes en los que se ha predicho que tendrán mal pronóstico podrían cambiarse a una terapia anticancerosa alternativa. Esto permite un mejor índice de supervivencia a largo plazo, ya que cuanto antes se utilice una terapia anticancerosa alternativa, más probable será que conduzca a un efecto beneficioso.

Continuando con la fig. 3, etapa 312, puede efectuarse el análisis del clasificador con ocultación. Esto significa que el algoritmo clasificador usa los espectros etiquetados por clase para la clasificación de los espectros de análisis (p. ej., de pacientes con cáncer nuevos) para determinar si el paciente con cáncer que tiene el mísmo cáncer que los pacientes con cáncer de los espectros etiquetados por clase responderá al fármaco anticanceroso. El clasificador puede generarse mediante el uso del clasificador VMPK probabilístico, tal como se describe de aquí en adelante. Como resultado del clasificador puede haber tres etiquetas de clase potenciales, "bueno", "malo" o "no definido". Una etiqueta de clase o clasificación de "bueno" significa que el clasificador, al procesar el espectro de análisis, determina que el espectro de análisis está en el mismo grupo que el grupo "bueno" de los espectros etiquetados por clase. Los resultados de dicho análisis con ocultación se muestran en la fig. 14, y confirman los resultados de la fase de desarrollo.

En la etapa 314 de la fig. 3, y como se ha descrito previamente, puede realizarse una visualización, en la que la visualización puede incluir herramientas para realizar (i) el promedio de los espectros, (ii) la variación espectral, y (iii) la localización de características. Estas herramientas de visualización pueden ser útiles con fines diagnósticos.

Si el clasificador determina que el espectro de análisis se relaciona más estrechamente con el grupo "bueno" de espectros, entonces los espectros de análisis se clasificarán como "buenos" y se podría prescribir al paciente el fármaco anticanceroso con un cierto nivel de confianza en que él o ella responderán. Si el clasificador determina que el espectro de análisis se relaciona más estrechamente con el grupo "malo" de espectros, entonces los espectros de análisis se clasificarán como "malos" y no se prescribirá al paciente el fármaco anticanceroso. Si no puede determinarse que el espectro de análisis esté asociado ni con el grupo "bueno" ni con el grupo "malo" de espectros etiquetados por clase, entonces el espectro de análisis se clasificará como "indeterminado" y no se prescribirá al paciente el fármaco anticanceroso.

La tabla VII presenta otro conjunto de valores de promedios de picos de diferenciación, similares a los de la tabla V, determinados mediante la extracción de características y los algoritmos de selección de las etapas 308 y 310 de la fig. 3. Estos espectros se clasifican y etiquetan mediante el clasificador de la etapa 312 de la fig. 3 como "buenos", "malos" o "no definidos". Tal como se enumera, los espectros "malos" tienen picos de diferenciación que tienen desviaciones típicas grandes, típicamente mayores que la amplitud del pico, de modo que el pico no puede determinarse. Los espectros clasificados como "buenos" tienen picos de diferenciación que tienden a tener amplitudes y desviaciones típicas menores. Los espectros "no definidos" están en cierto modo en el medio, con amplitudes de los picos de diferenciación que son pequeñas a lo largo de ciertas localizaciones m/z y mayores a lo largo de otras.

TABLA VII Picos de diferenciación y desviaciones típicas de ejemplo

Grupo/MZ	5794,38	5868,02	11483,44	11572,81	11729,95	12495,04
Malo	190,83 ± 207,43	232,88 ± 301,35	798,03 ± 964,81	1451,46 ± 1541,45	1747,09 ± 2208,33	97,96 ± 109,81
Bueno	8,74 ± 3,69	6,40 ± 4,34	6,06 ± 6,21	15,34 ± 17,77	20,15 ± 19,91	2,68 ± 4,50
No definido	17,62 ± 6,76	16,62 ± 7,94	37,84 ± 20,51	89,82 ± 47,87	105,52 ± 53,71	8,18 ± 6,08

El nivel de confianza se basa en la probabilidad de asociación con el conjunto de espectros de prácticas ajustada mediante el parámetro delta-p para el algoritmo VMPK probabilístico. El parámetro delta-p puede incrementarse o reducirse dependiendo del nivel de confianza deseado para asociar espectros de análisis con el conjunto de prácticas. En un estudio con análisis con ocultación, el parámetro delta-p se ajustó a 0,2 y se obtuvo como resultado un resultado de predicción con una precisión del 92 %.

Aunque la fig. 11 es útil para representar gráficamente los espectros en el espacio de características bidimensional, los espectros del mundo real dan típicamente un resultado de un espacio de características de 8-12 dimensiones, alcanzando a menudo 8-12 dimensiones o más. Puede determinarse que un espacio de características con más o menos dimensiones es adecuado o necesario para determinar si un paciente con cáncer responderá a un fármaco anticanceroso. Por tanto, en ciertas realizaciones, un facultativo podría utilizar sólo uno o dos picos diferenciales, en otras realizaciones se usarían tres o cuatro picos diferenciales, en otras realizaciones más se usarían cinco o seis picos diferenciales, aún en otras realizaciones más se usarían siete u ocho picos diferenciales, y aún en otras realizaciones más se usarían nueve o diez picos diferenciales y en otras realizaciones se usarían once o doce picos diferenciales. De hecho, la invención prevé la adición incluso de más de doce picos diferenciales. La determinación del número de características que proporcionan suficiente información para ser determinista puede basarse en una serie de factores, que incluyen la amplitud de las características, la clasificación de los espectros, y la respuesta del paciente al tratamiento anticanceroso, por ejemplo.

Continuando con la fig. 3, puede utilizarse una base de datos, tal como la base de datos 220 (fig. 2), para recibir y almacenar picos de diferenciación, diagnósticos del espectrómetro de masas, y/u otros parámetros resultantes del proceso de clasificación y diagnóstico, tal como se ha descrito. Estos parámetros pueden almacenarse y usarse para la futura clasificación de espectros nuevos de pacientes con cáncer nuevos. Al final, la base de datos puede poblarse hasta el punto de que la precisión y la fiabilidad al clasificar los espectros de análisis aseguren de forma sustancial con una elevada probabilidad, tal como del 98 %, que un paciente con cáncer responderá al fármaco anticanceroso.

La fig. 12 es una gráfica de Kaplan-Meier 1200 de los datos de análisis que muestra los índices de supervivencia de los grupos de pacientes clasificados de acuerdo con los principios de la presente invención. La gráfica de Kaplan-Meier 1200 es una gráfica de mortalidad que indica los índices de supervivencia a lo largo de ciertas duraciones de tiempo. Tal como se muestra, los pacientes con cáncer que se categorizaron como "buenos" vivieron durante más largo tiempo, debido a que habían recibido el fármaco anticanceroso. Los pacientes con cáncer que se categorizaron como "malos" sufrieron un descenso pronunciado en los primeros pocos meses. Los pacientes con cáncer que se categorizaron como "no definidos" disminuyeron de forma constante, con un bajo índice de supervivencia. Este gráfico se obtuvo en la fase de descubrimiento probando ensayando en las muestras japonesas 1 y 2 un clasificador entrenado en las muestras italianas 1.

La fig. 13 es una gráfica de Kaplan-Meier 1300 similar a la fig. 12 en la que un clasificador entrenado en las muestras japonesas 1 y 2, se ensayó en el conjunto italiano 1. Tal como se muestra, se predijo que los pacientes cuyos espectros asociados se clasificaron como "buenos" tendrían una vida prolongada a partir del tratamiento con el fármaco anticanceroso. Se predijo que los pacientes clasificados como "malos" tendrían un índice de mortalidad pronunciado con un pequeño porcentaje que llegaría más allá de un año. Los pacientes clasificados como "no definidos" tuvieron un descenso constante y no se predijo que ninguno viviera más allá de seis meses. Con los análisis clínicos se demostró que estas predicciones eran precisas.

La fig. 14 es una gráfica de Kaplan-Meier 1400 similar a las fig. 12 y 13 obtenidas a partir del uso con ocultación del clasificador validado sobre las muestras italianas 2. En el momento del análisis, no se tenía conocimiento de los datos de supervivencia, ya que se mantuvieron de forma confidencial. Después de realizar la clasificación, los datos de supervivencia se desvelaron, y las curvas de la fig. 14 confirmaron los resultados del análisis del desarrollo. Tal como se muestra, se predijo que los pacientes clasificados como "buenos" tendrían un índice de supervivencia prolongado, y los clasificados como "malos" tuvieron una reducción pronunciada con una esperanza de vida más limitada. En este caso concreto, el análisis se llevó a cabo con un delta-p bajo, para que no hubiese pacientes clasificados como "no definidos", De nuevo, los resultados fueron coherentes con los análisis clínicos reales.

La fig. 15 es un diagrama por bloques de un proceso de ejemplo 1500 para determinar si un paciente con cáncer responderá a un fármaco anticanceroso de acuerdo con los principios de la presente invención. El proceso 1500 empieza en la etapa 1502, en la que se obtiene un espectro de análisis producido mediante un espectrofotómetro de masas a partir de un suero producido a partir de un paciente con cáncer. En la etapa 1504, El espectro de análisis se procesa para determinar una relación con un grupo de espectros etiquetados por clase producidos a partir de los sueros respectivos de otros pacientes en el mismo o similar estadio clínico del cáncer y que se sabe que han respondido o no han respondido al fármaco. La relación significa que es más probable que el espectro de análisis esté asociado o tenga las mismas o similares características que las de unos u otros espectros etiquetados por clase. El fármaco anticanceroso puede ser uno que trate el carcinoma no microcítico de pulmón. En la etapa 1506, una determinación, basada en la relación del espectro de análisis con el grupo del espectro clasificado, si el paciente responderá al fármaco anticanceroso. Responder significa que el fármaco anticanceroso tendrá algún beneficio positivo para el paciente con cáncer. La respuesta positiva prolongara de forma esperanzadora la vida del paciente, pero que se trate al paciente con el fármaco anticanceroso puede dar como resultado otros beneficios positivos.

Los biomarcadores determinados mediante la presente invención pueden ser cualquier tipo de parámetros cuantificables que aparezcan como un pico en una espectroscopia de masas de un espectro. El parámetro que causa el pico de la espectroscopia de masas puede estar causado por cualquier sustancia, incluyendo pero no limitándose a, enzimas específicas, hormonas, ARNm, ADN, ARN, proteínas, lípidos, vitaminas, minerales,

metabolitos, y compuestos químicos. Además, los biomarcadores pueden determinarse a partir de cualquier tejido o fluido recogido del paciente, incluyendo pero no limitándose a, suero, glóbulos rojos, glóbulos blancos, uñas, piel, pelo, tejido biopsiado, líquido cefalorraquídeo, médula ósea, orina, heces, esputo, bilis, líquido broncoalveolar, líquido pleural, y líquido peritoneal. Los biomarcadores pueden reflejar una diversidad de características de la enfermedad, incluyendo el nivel de exposición a un desencadenante ambiental o genético, un elemento del propio proceso de la enfermedad, una etapa intermedia entre la exposición y el comienzo de la enfermedad, o un factor independiente asociado con el estado de la enfermedad, pero no causante de la patogenia. Por definición, se prevé que los principios de la presente invención también pueden ser aplicables para determinar estadios específicos de enfermedad y trastornos.

REIVINDICACIONES

1. Un método para determinar si un paciente con carcinoma no microcítico de pulmón responderá a un fármaco, comprendiendo el método:

a. obtención de un espectro de análisis producido por un espectrómetro de masas a partir de un suero producido a partir del paciente;
b. procesamiento del espectro de análisis para determinar una relación con un grupo de espectros etiquetados por clase producidos a partir del suero respectivo de otros pacientes que tienen el mismo o similar estadio clínico de la enfermedad y que se sabe que han respondido o no han respondido al fármaco; y
c. determinación, basada en la relación del espectro de análisis con el grupo de espectros etiquetados por clase, de si el paciente responderá al fármaco, en el que el fármaco es gefitinib o erlotinib.

2. Un método de acuerdo con la reivindicación 1, en el que los espectros etiquetados por clase comprenden una pluralidad de picos de diferenciación, los cuales son utilizables para determinar si es o no probable que el paciente responda al fármaco.

3. Un método de acuerdo con la reivindicación 2, en el que uno o más de los picos de diferenciación son como se define en la tabla IV:

mz central	mz bajo	mz alto	anchura = mz alto-mz bajo
5763,791	5732,131	5795,45	63,3
5843,241	5811,097	5875,384	64,3
6433,973	6398,186	6469,759	71,6
11445,75	11376,15	11515,34	139,2
11529,52	11459,32	11599,73	140,4
11685,37	11614,03	11756,71	142,7
11759,16	11687,28	11831,04	143,8
11903,24	11830,3	11976,18	145,9
12452,38	12375,37	12529,4	154
23354,35	23183,57	23525,13	341,6
23451,25	23279,53	23622,97	343,4
66702,45	65902,02	67502,88	1600

4. El método de acuerdo con la reivindicación 1, en el que el procesamiento incluye la selección de al menos ocho picos del espectro de análisis para determinar la relación del espectro de análisis y el grupo de espectros etiquetados por clase para permitir la determinación de si el paciente responderá al fármaco anticanceroso.

5. El método de acuerdo con la reivindicación 1, en el que la obtención del espectro de análisis se obtiene a partir de un espectrómetro de masas de desorción/ionización láser asistida por matriz (MALDI).

6. El método de acuerdo con la reivindicación 1, que comprende, además, el preprocesamiento del espectro de análisis antes del procesamiento para preparar el espectro de análisis de acuerdo con el procesamiento realizado en el grupo de espectros etiquetados por clase.

7. El método de acuerdo con la reivindicación 6, en el que el preprocesamiento incluye la reducción del fondo contenido en el espectro de análisis.

8. El método de acuerdo con la reivindicación 7, en el que el preprocesamiento incluye la normalización del espectro de análisis con el fondo reducido.

9. El método de acuerdo con la reivindicación 8, en el que el preprocesamiento incluye además la selección de los picos del espectro de análisis normalizado con el fondo reducido.

10. El método de acuerdo con la reivindicación 9, en el que el preprocesamiento incluye además el alineamiento espectral de los picos seleccionados del espectro de análisis normalizado con el fondo reducido.

11. El método de acuerdo con la reivindicación 1, en el que la determinación incluye la determinación de si el paciente responderá al fármaco gefitinib.

12. El método de acuerdo con la reivindicación 1, en el que el procesamiento del espectro de análisis incluye la selección de una pluralidad de picos de diferenciación del espectro de análisis para su procesamiento en relación con los picos del grupo de espectros etiquetados por clase.

FIG. 1

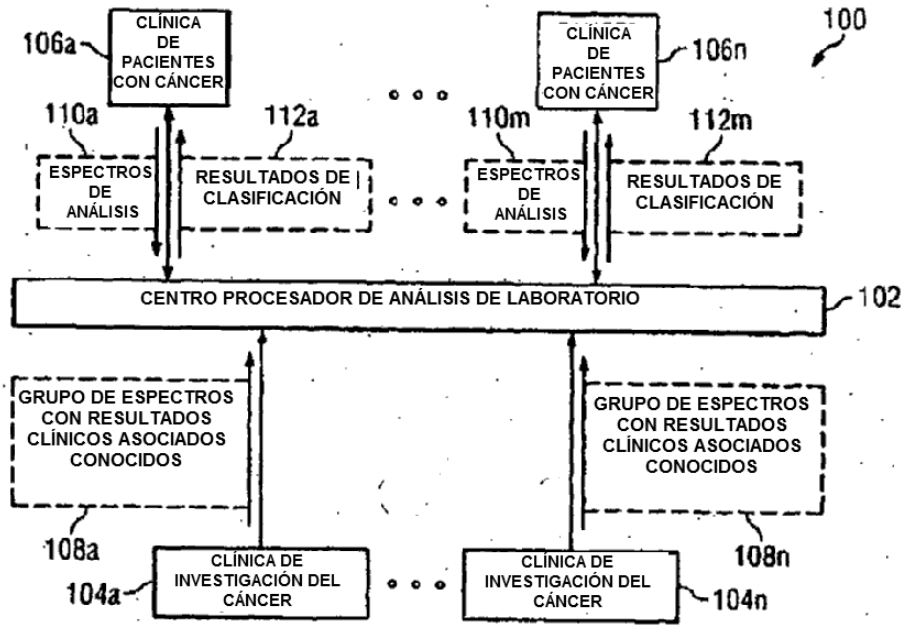

100
106a
CLÍNICA DE PACIENTES CON CÁNCER
106n
CLÍNICA DE PACIENTES CON CÁNCER
110a
112a
110m
112m
ESPECTROS DE ANÁLISIS
RESULTADOS DE CLASIFICACIÓN
ESPECTROS DE ANÁLISIS
RESULTADOS DE CLASIFICACIÓN
CENTRO PROCESADOR DE ANÁLISIS DE LABORATORIO
102
GRUPO DE ESPECTROS CON RESULTADOS CLÍNICOS ASOCIADOS CONOCIDOS
GRUPO DE ESPECTROS CON RESULTADOS CLÍNICOS ASOCIADOS CONOCIDOS
108a
108n
104a
CLÍNICA DE INVESTIGACIÓN DEL CÁNCER
CLÍNICA DE INVESTIGACIÓN DEL CÁNCER
104n

FIG. 2

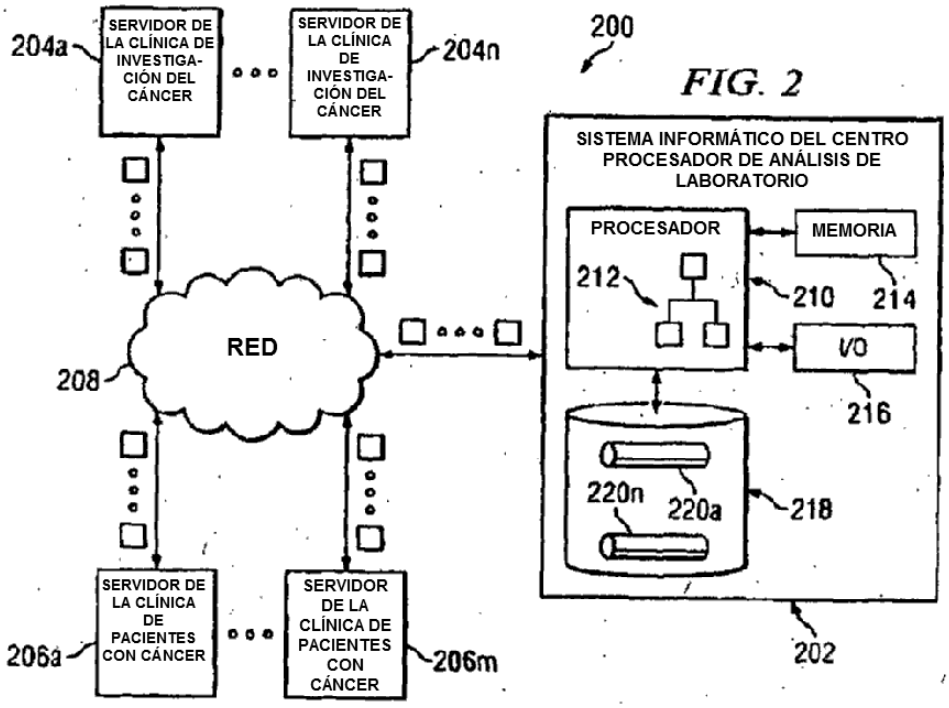

200
204a
SERVIDOR DE LA CLÍNICA DE INVESTIGACIÓN DEL CÁNCER
SERVIDOR DE LA CLÍNICA DE INVESTIGACIÓN DEL CÁNCER
204n
SISTEMA INFORMÁTICO DEL CENTRO PROCESADOR DE ANÁLISIS DE LABORATORIO
PROCESADOR
MEMORIA
212
210
214
I/O
216
RED
208
220n
220a
218
SERVIDOR DE LA CLÍNICA DE PACIENTES CON CÁNCER
SERVIDOR DE LA CLÍNICA DE PACIENTES CON CÁNCER
206a
206m
202

FIG. 3

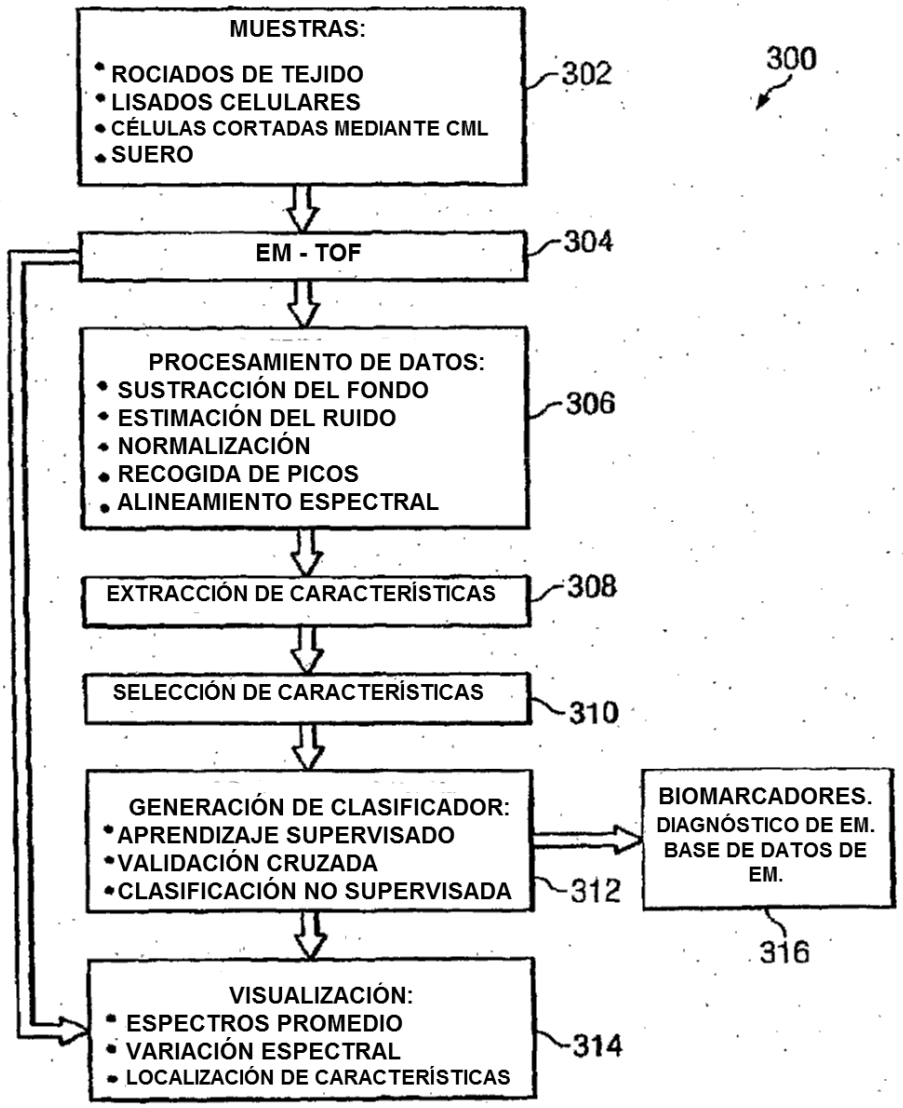
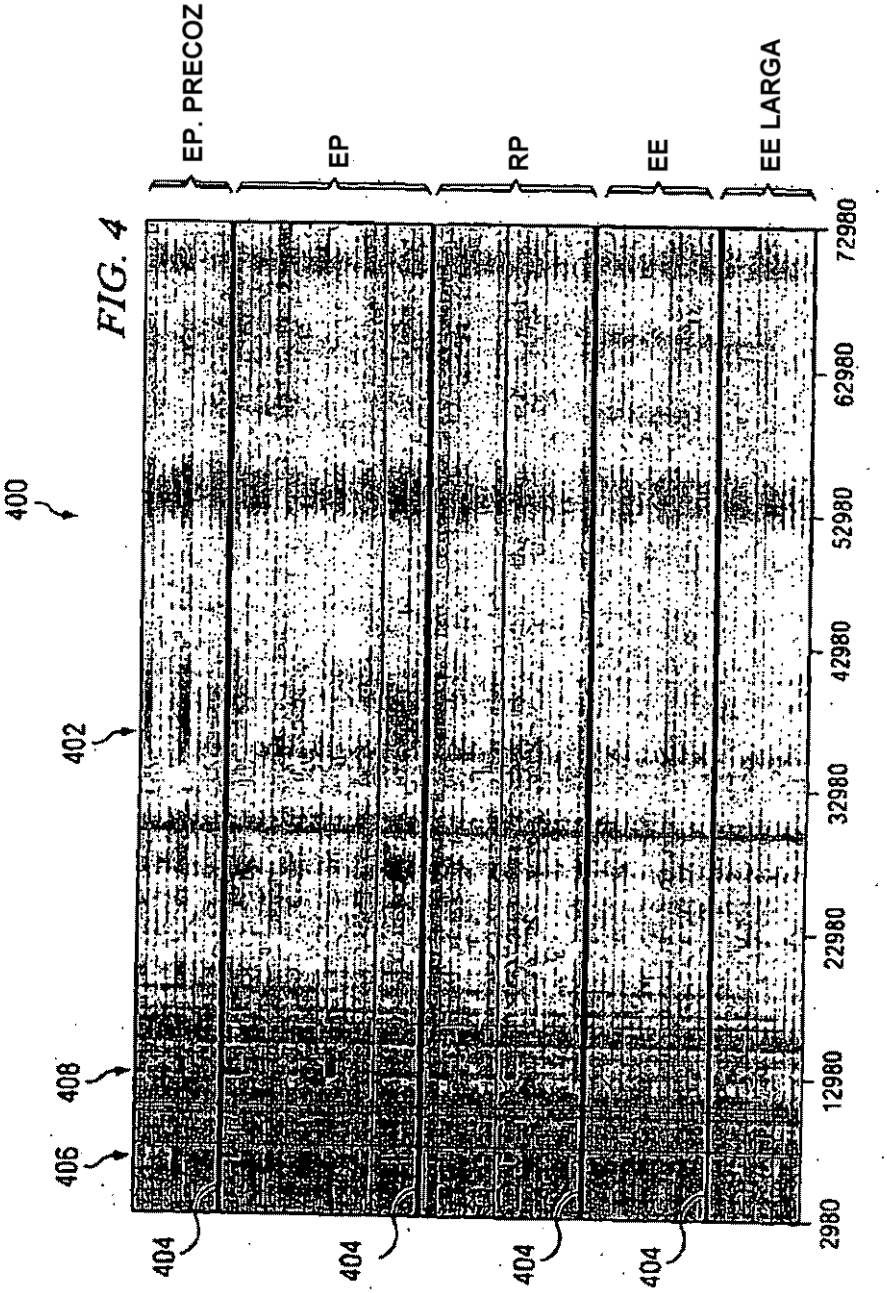

MUESTRAS:
• ROCIADOS DE TEJIDO
• LISADOS CELULARES
• CÉLULAS CORTADAS MEDIANTE CML
• SUERO
302
300
EM - TOF
304
PROCESAMIENTO DE DATOS:
• SUSTRACCIÓN DEL FONDO
• ESTIMACIÓN DEL RUIDO
• NORMALIZACIÓN
• RECOGIDA DE PICOS
• ALINEAMIENTO ESPECTRAL
306
EXTRACCIÓN DE CARACTERÍSTICAS
308
SELECCIÓN DE CARACTERÍSTICAS
310
GENERACIÓN DE CLASIFICADOR:
• APRENDIZAJE SUPERVISADO
• VALIDACIÓN CRUZADA
• CLASIFICACIÓN NO SUPERVISADA
312
BIOMARCADORES.
DIAGNÓSTICO DE EM.
BASE DE DATOS DE EM.
316
VISUALIZACIÓN:
• ESPECTROS PROMEDIO
• VARIACIÓN ESPECTRAL
• LOCALIZACIÓN DE CARACTERÍSTICAS
314

FIG. 4


EP. PRECOZ
EP
RP
EE
EE LARGA
400
402
404
406
408
72980
62980
52980
42980
32980
22980
12980
2980

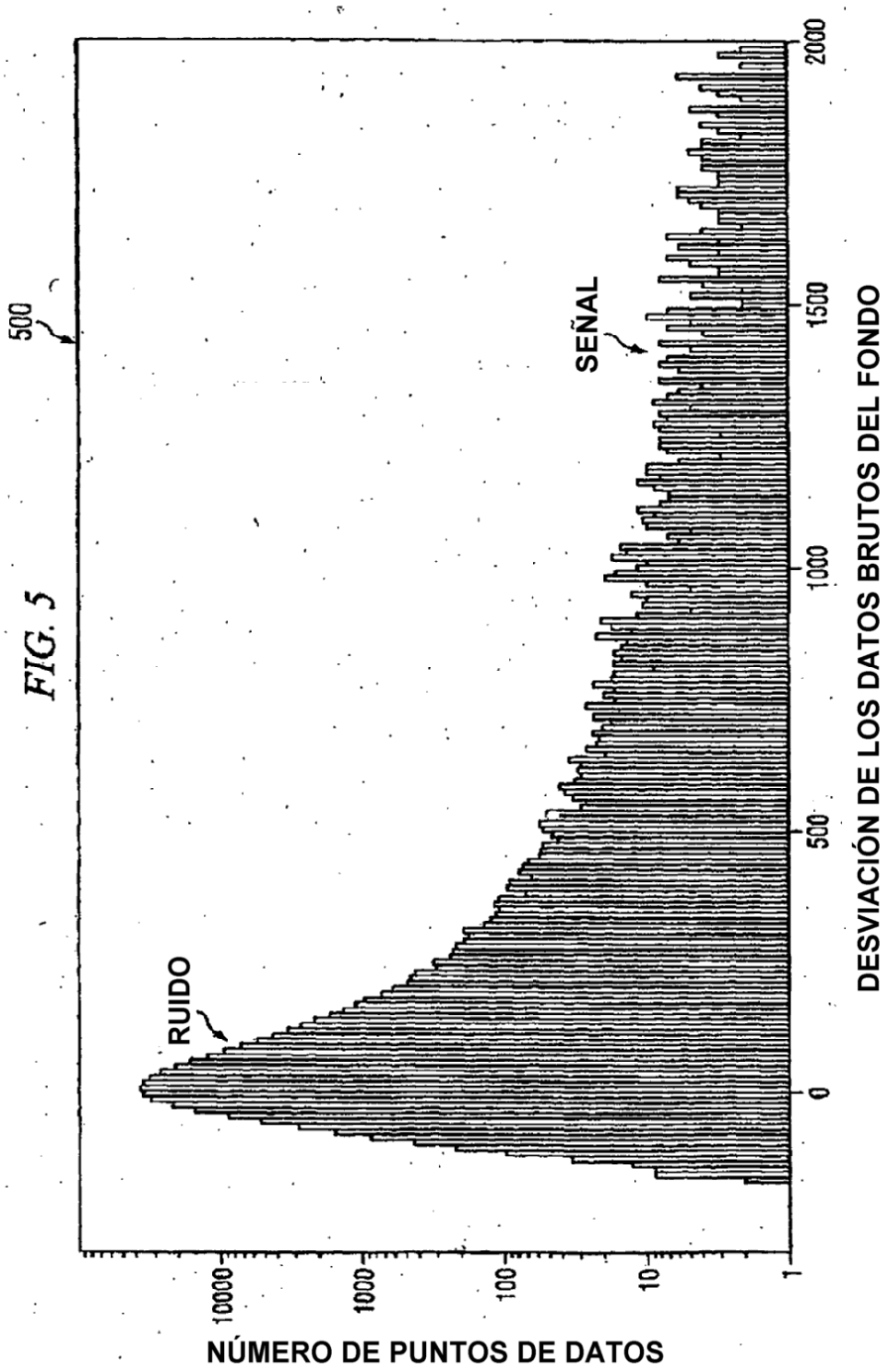

500
RUIDO
SEÑAL
NÚMERO DE PUNTOS DE DATOS
10000
1000
100
10
1
0
500
1000
1500
2000
DESVIACIÓN DE LOS DATOS BRUTOS DEL FONDO

FIG. 5

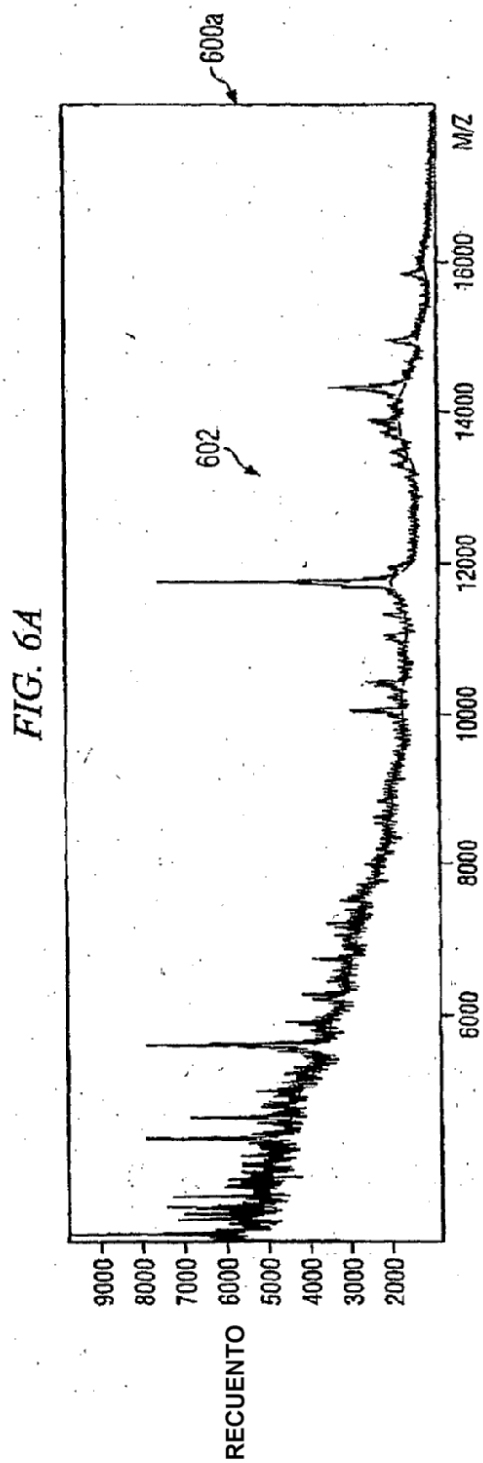

600a
602
RECUENTO
9000
8000
7000
6000
5000
4000
3000
2000
6000
8000
10000
12000
14000
16000
M/Z

FIG. 6A

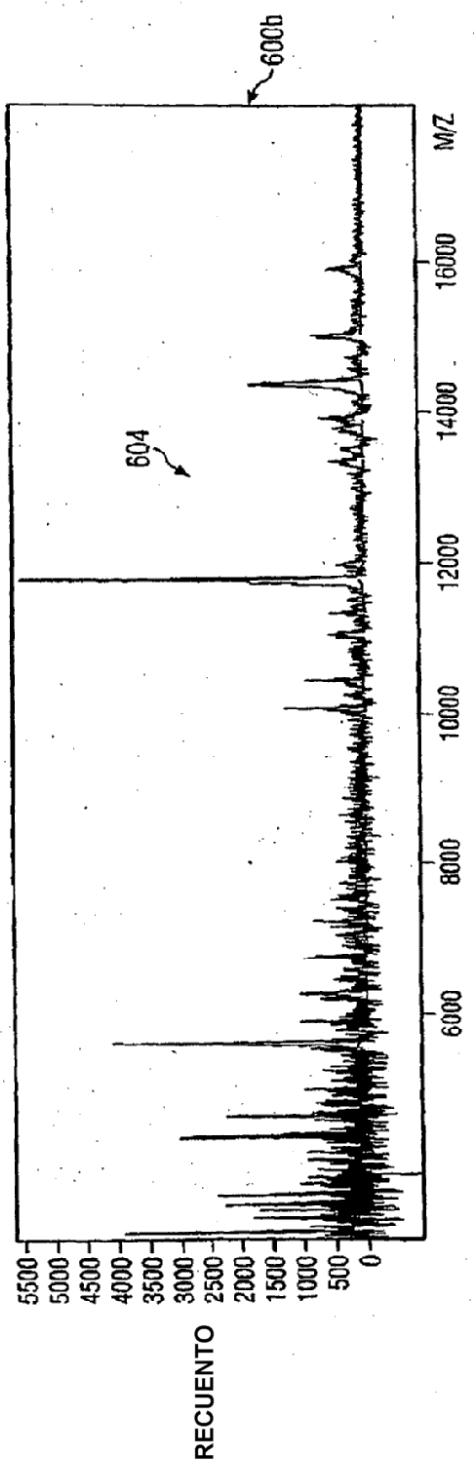

600b
604
RECUENTO
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
6000
8000
10000
12000
14000
16000
M/Z

FIG. 6B

FIG. 7A

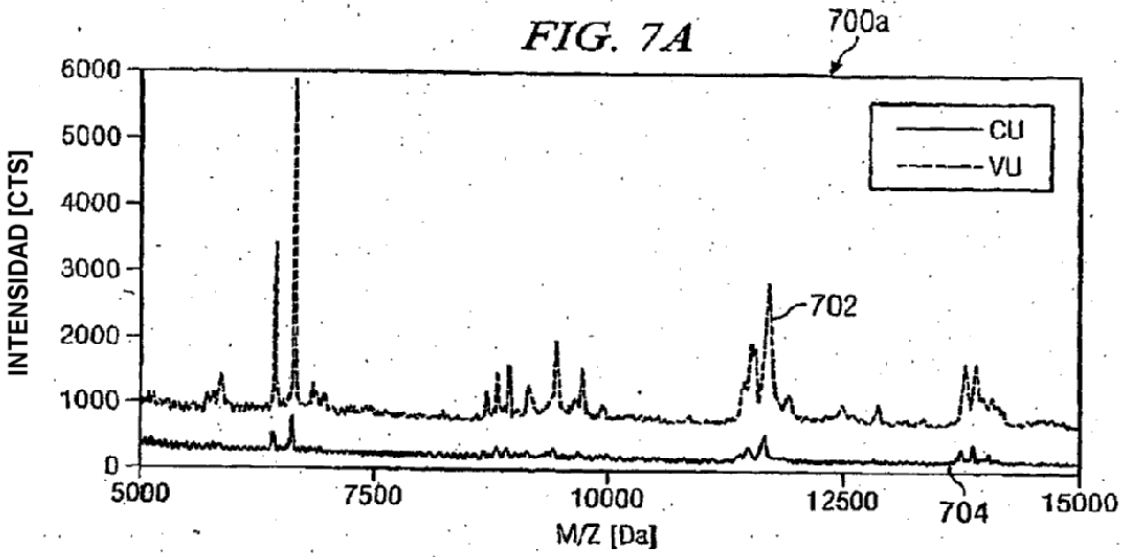
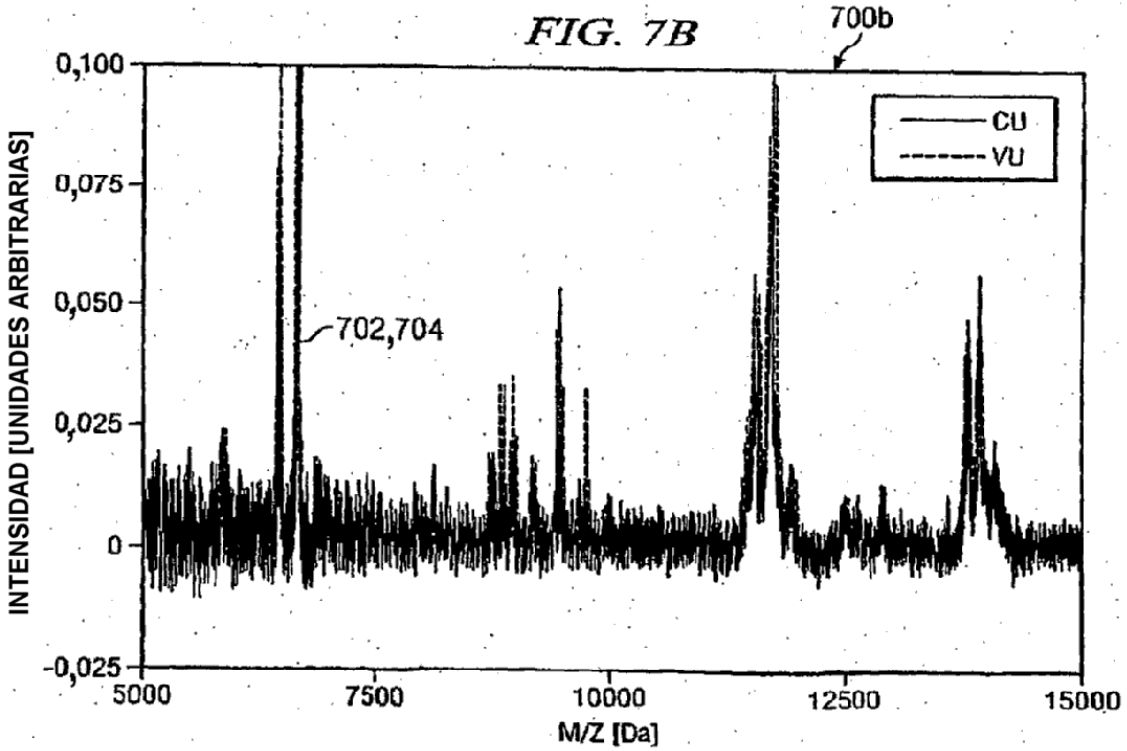

700a
6000
5000
4000
3000
2000
1000
0
INTENSIDAD [CTS]
CU
VU
702
704
5000
7500
10000
12500
15000
M/Z [Da]

FIG. 7B


700b
0,100
0,075
0,050
0,025
0
-0,025
INTENSIDAD [UNIDADES ARBITRARIAS]
CU
VU
702,704
5000
7500
10000
12500
15000
M/Z [Da]

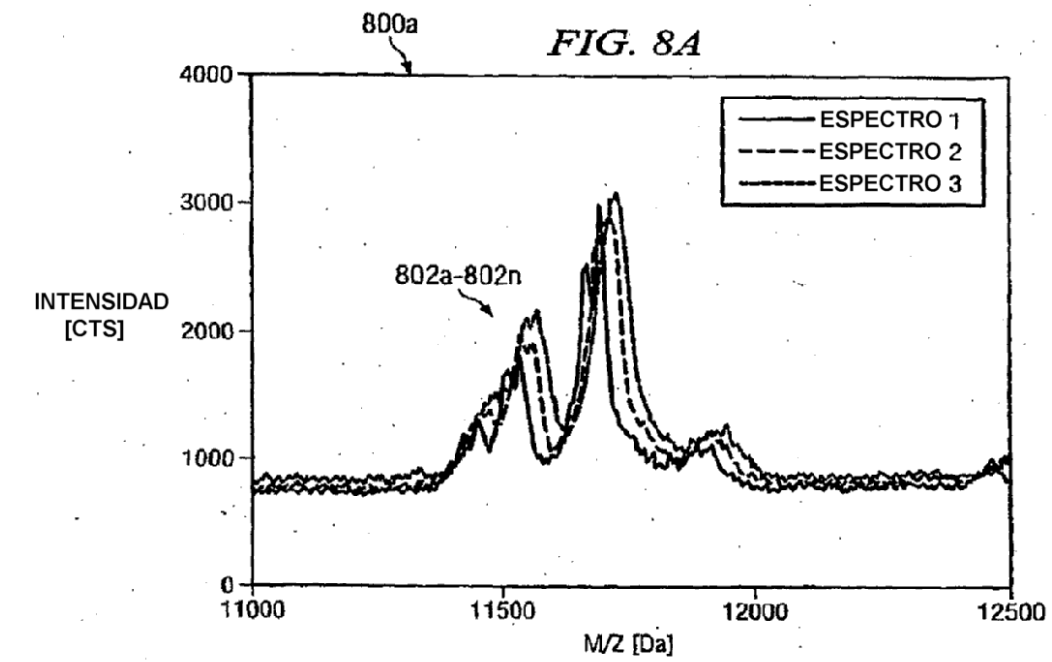

800a
4000
3000
2000
1000
0
INTENSIDAD [CTS]
802a-802n
ESPECTRO 1
ESPECTRO 2
ESPECTRO 3
11000
11500
12000
12500
M/Z [Da]

FIG. 8A

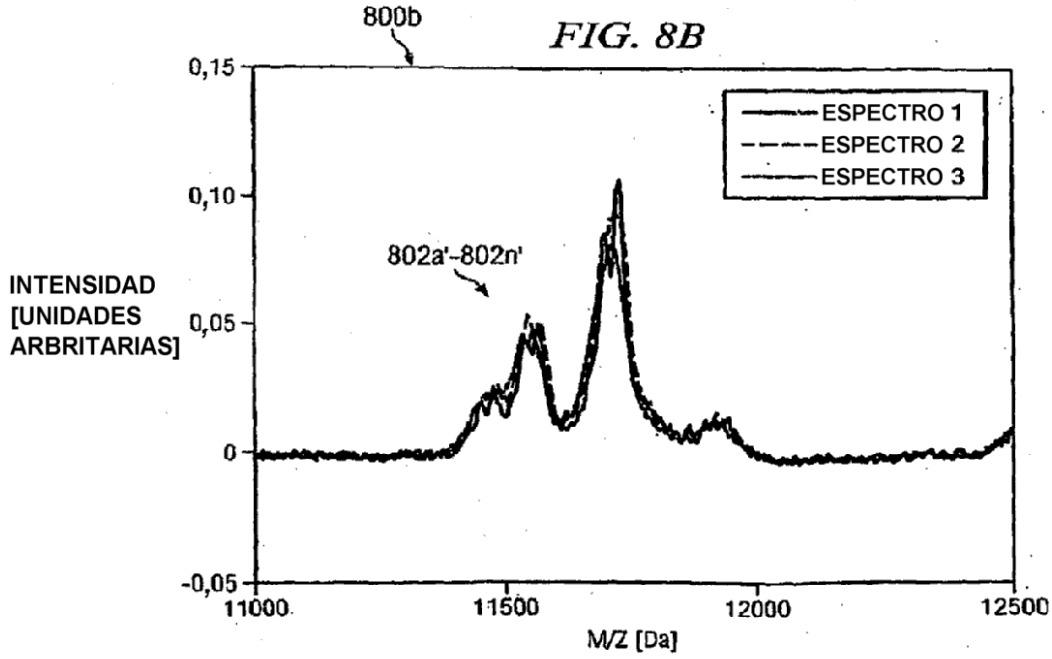

800b
0,15
0,10
0,05
0
-0,05
INTENSIDAD [UNIDADES ARBRITARIAS]
802a'-802n'
ESPECTRO 1
ESPECTRO 2
ESPECTRO 3
11000
11500
12000
12500
M/Z [Da]

FIG. 8B

FIG. 9

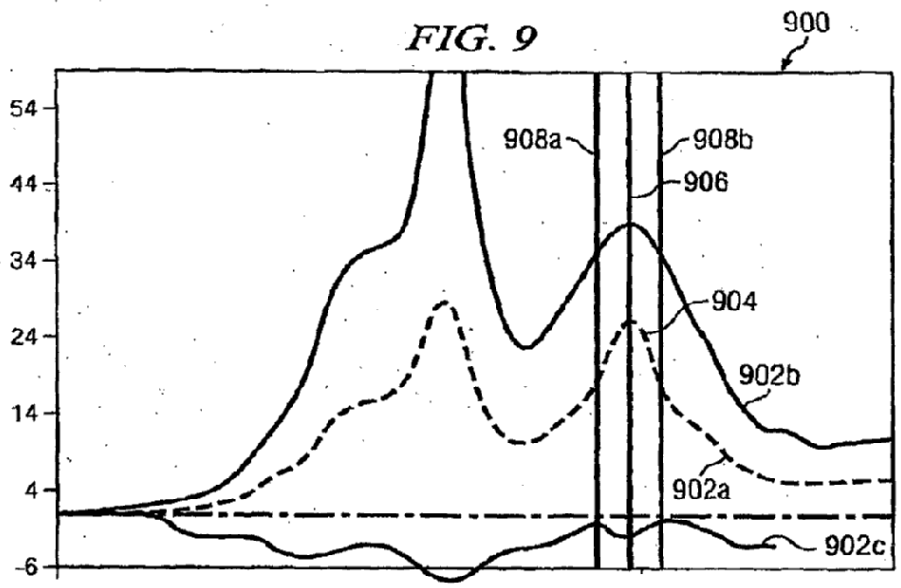

900
54
44
34
24
14
4
-6
908a
908b
906
904
902b
902a
902c

FIG. 10A

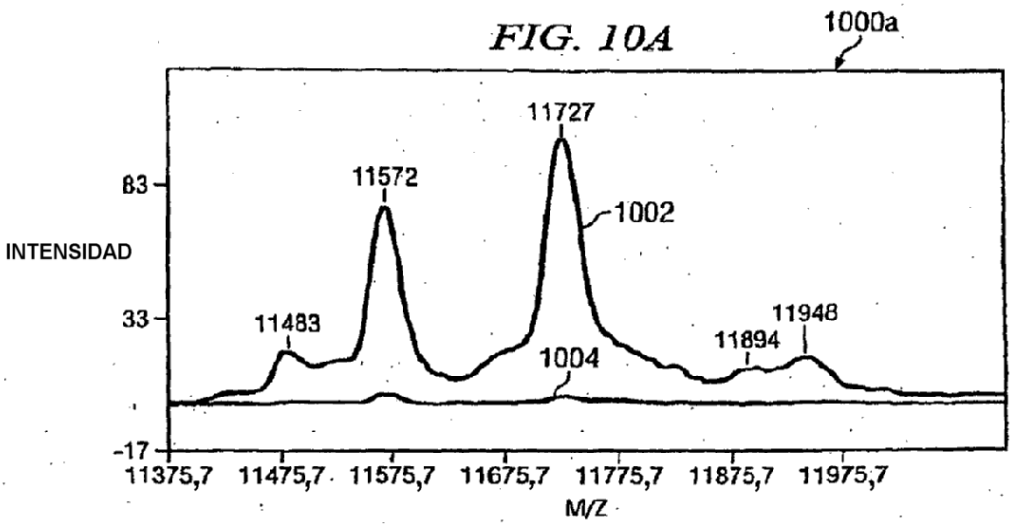

1000a
INTENSIDAD
83
33
-17
11727
11572
1002
11483
11948
11894
1004
11375,7
11475,7
11575,7
11675,7
11775,7
11875,7
11975,7
M/Z

FIG. 10B

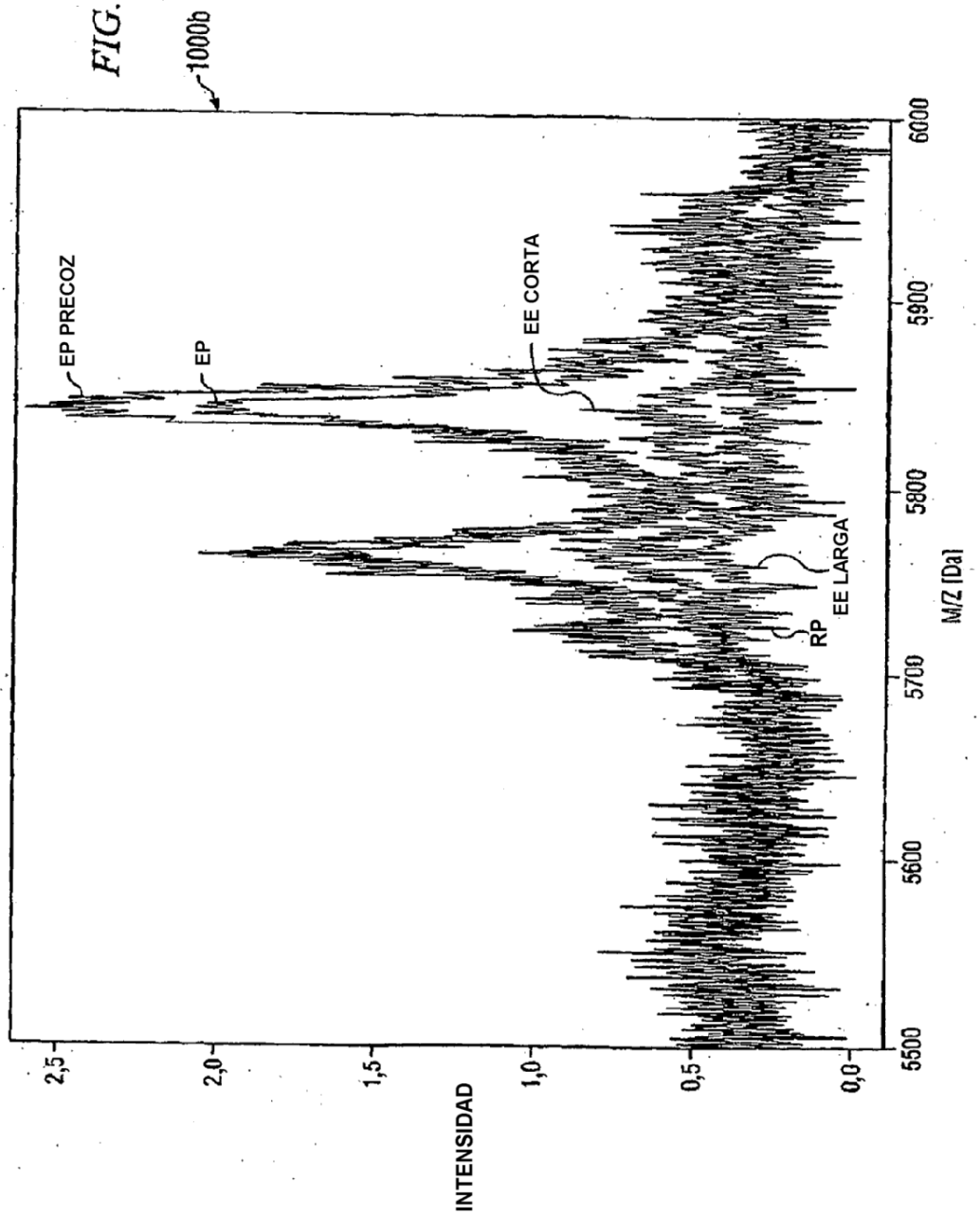

1000b
EP PRECOZ
EP
EE CORTA
RP
EE LARGA
INTENSIDAD
2,5
2,0
1,5
1,0
0,5
0,0
5500
5600
5700
5800
5900
6000
M/Z [Da]

FIG. 10C

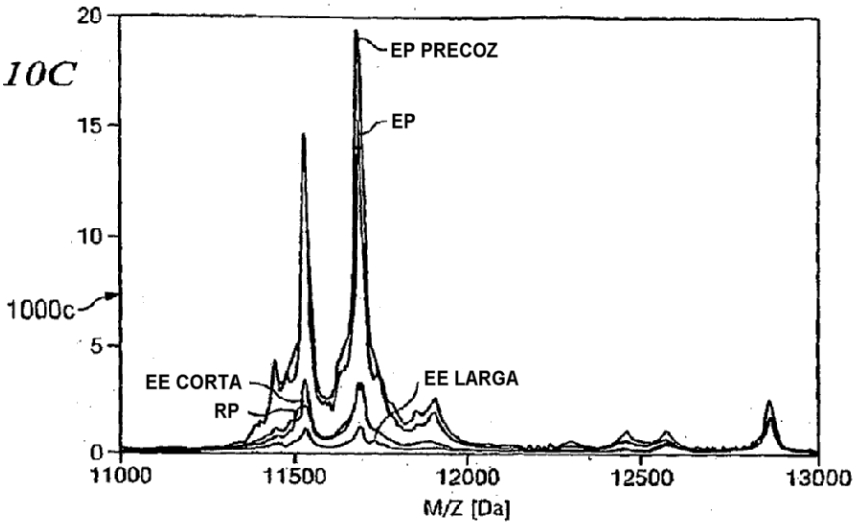


FIG. 11

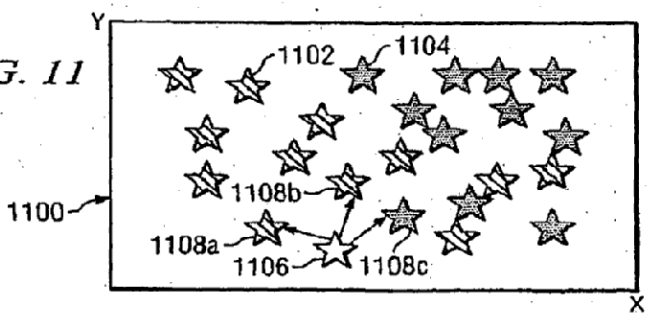
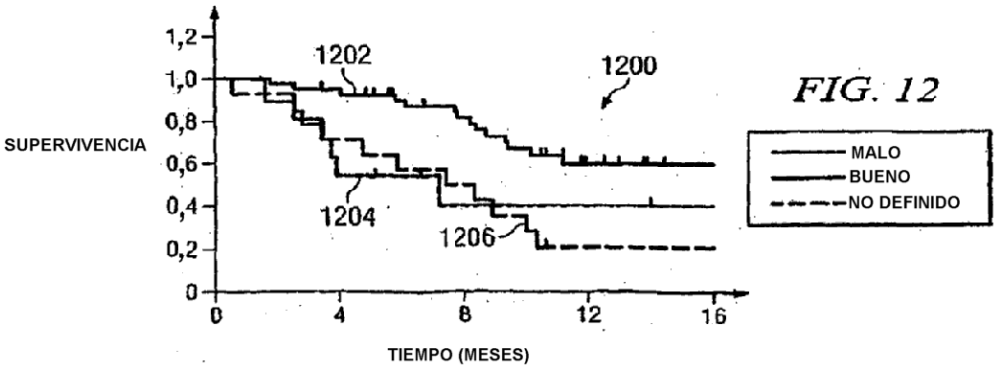


FIG. 12



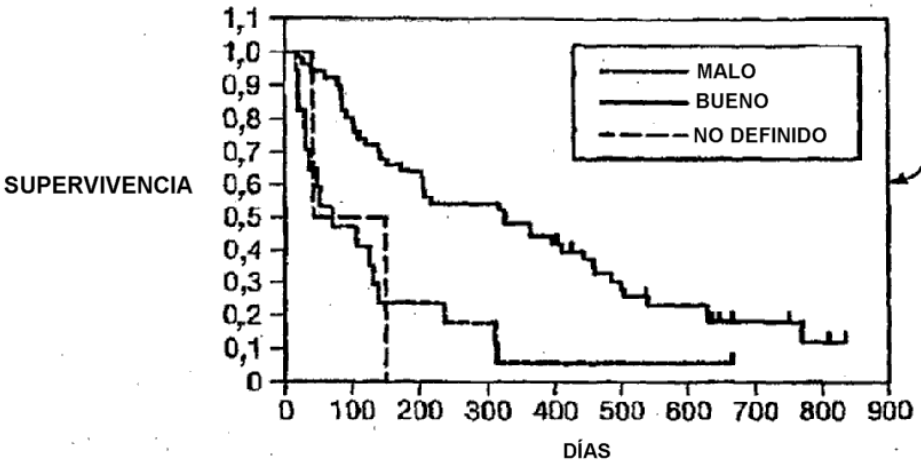


FIG. 13

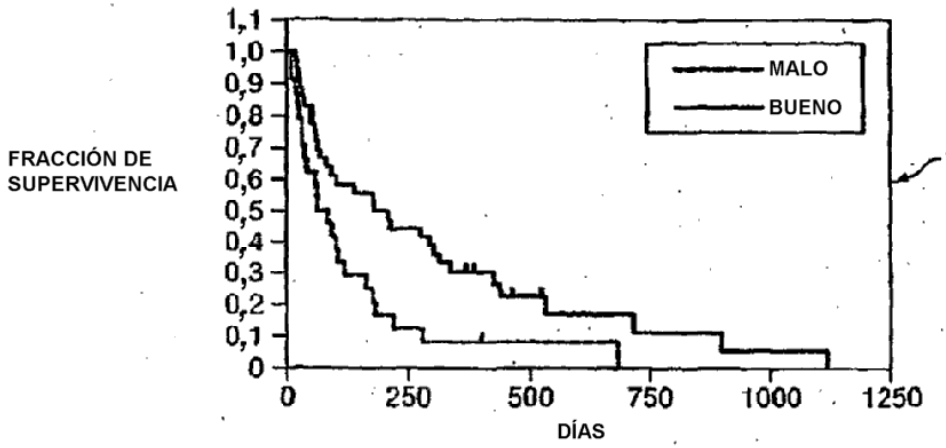


FIG. 14

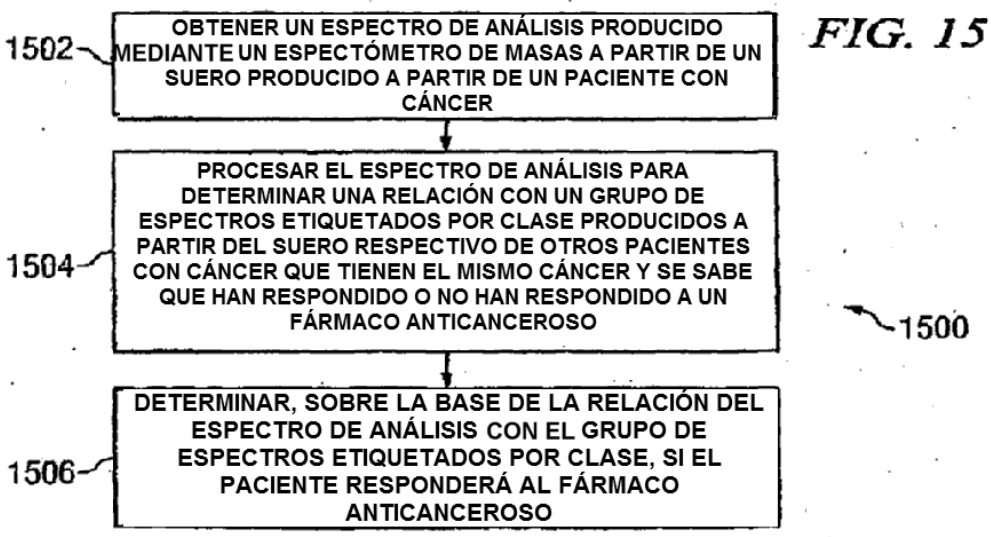


FIG. 15