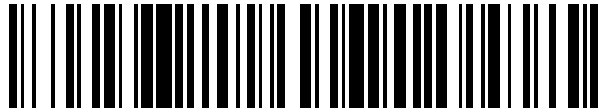


19



OFICINA ESPAÑOLA DE  
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 665 551**

51 Int. Cl.:

**G06K 9/00** (2006.01)

**G06K 9/62** (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **30.11.2012 PCT/IB2012/056859**

87 Fecha y número de publicación internacional: **06.06.2013 WO13080169**

96 Fecha de presentación y número de la solicitud europea: **30.11.2012 E 12798883 (0)**

97 Fecha y número de publicación de la concesión europea: **10.01.2018 EP 2798575**

54 Título: **Procedimiento de identificación de microorganismos por espectrometría de masas y normalización de resultados**

30 Prioridad:

**02.12.2011 EP 11306609**  
**02.12.2011 US 201161566029 P**

45 Fecha de publicación y mención en BOPI de la traducción de la patente:  
**26.04.2018**

73 Titular/es:

**BIOMÉRIEUX INC. (100.0%)**  
**100 Rodolphe Street**  
**Durham, North Carolina 27712, US**

72 Inventor/es:

**STRUBEL, GRÉGORY;**  
**ARSAC, MAUD;**  
**DESSEREE, DENIS;**  
**COTTE-PATTAT, PIERRE-JEAN y**  
**MAHE, PIERRE**

74 Agente/Representante:

**LEHMANN NOVO, María Isabel**

**ES 2 665 551 T3**

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

## DESCRIPCIÓN

Procedimiento de identificación de microorganismos por espectrometría de masas y normalización de resultados

## 5 Campo de la invención

La presente invención se refiere a la identificación de microorganismos, en particular de bacterias, mediante la espectrometría de masas.

## 10 Estado de la técnica

Se conoce utilizar la espectrometría de masas para identificar microorganismos, y más particularmente bacterias. Se prepara una muestra del microorganismo y después se adquiere y se pretrata un espectro de masas de la muestra, en particular para eliminar la línea de base (comúnmente denominada "baseline") y para eliminar el ruido. Se detectan entonces los picos del espectro pretratado y la lista de picos así obtenida se "analiza" y se "compara" con la ayuda de herramienta de clasificación con datos de una base de conocimientos construida a partir de listas de picos, cada una asociada a un microorganismo o a un grupo de microorganismos (cepa, clasificación, familia, etc.) identificado.

20 Entre las herramientas de clasificaciones, se conocen las clasificaciones SVM (por "Support Vector Machine") del tipo (uno contra todos) (a continuación, clasificación "SVM-UCT"). Una clasificación SVM del tipo "uno contra todos" consiste en determinar, para cada clase de objetos de un conjunto de clases, una frontera orientada que separa esta clase de las otras clases del conjunto. Se obtiene por lo tanto tantas clases "uno contra todos" como clases existen en el conjunto. La identificación de un objeto desconocido consiste entonces en consultar cada uno de los clasificadores calculando la distancia algébrica entre el objeto desconocido y la frontera asociada al clasificador. Habitualmente, el objeto desconocido se determina como perteneciente a la clasificación asociada a la mayor distancia calculada.

30 Este principio se ilustra brevemente en las figuras 1 y 2 que ilustran el caso muy simple de 3 microorganismos identificables mediante solo dos picos en un espectro de masas, por ejemplo los dos picos de mayor intensidad en el espectro de masas de los 3 microorganismos. El primer microorganismo se caracteriza por un primer pico localizado sobre un valor  $m_{11}$  y un segundo pico localizado en un valor  $m_{12}$  (figura 1A), el segundo microorganismo se caracteriza por un primer pico localizado sobre un valor  $m_{21}$  y un segundo pico localizado sobre un valor  $m_{22}$  (figura 1B), y el tercer microorganismo se caracteriza por un primer pico localizado en un valor  $m_{31}$  y un segundo pico localizado en un valor  $m_{32}$  (figura 1C).

La clasificación de tipo SVM-UTC consiste, en primer lugar, en adquirir un conjunto de espectros de masas de aprendizaje de cada uno de los microorganismos y determinar la localización de los dos picos implicados en cada

40 espectro adquirido a fin de formar un conjunto de vectores  $\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$  de aprendizaje, siendo  $P_1$  la posición medida del primer pico, y siendo  $p_2$  la posición medida del segundo pico. Debido a la incertidumbre sobre la medición, se observa una dispersión de los valores de los vectores. En una segunda etapa, se calcula para cada microorganismo

45 una frontera que separa el conjunto de los vectores  $\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$  asociados al microorganismo, de los vectores  $\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$  asociados a los dos otros microorganismos. Se obtiene así Tres fronteras  $F_1$ ,  $F_2$  y  $F_3$  se obtiene así, tales como se representan en la figura 2, y están provistas de una orientación, por ejemplo la representada por las flechas en líneas de puntos.

La identificación de un microorganismo desconocido consiste entonces en adquirir uno o varios espectros de masa de

microorganismo, deducir un vector  $M$  de picos  $\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$  medido, y calcular la distancia algébrica, también denominada "margen" de este vector  $M$  a cada una de las fronteras orientadas  $F_1$ ,  $F_2$  y  $F_3$ . Se obtiene así un vector de distancias

50 algébricas, por ejemplo igual a  $\begin{pmatrix} -0,4 \\ +0,3 \\ -1,3 \end{pmatrix}$ . En el caso muy simple ilustrado, se podría por lo tanto deducir que el microorganismo desconocido es el segundo microorganismo.

Por supuesto, el caso ilustrado aquí es extremadamente simple. En la realidad, se trata de identificar un microorganismo entre centenares de microorganismos con un número de picos retenidos para la identificación que puede superar sustancialmente 1000 picos. Además, el caso ilustrado es también simple ya que los

microorganismos están muy alejados los unos de los otros y que la medición se ha realizado con suficientemente precisión para poder deducir de las distancias una información pertinente.

5 En los casos reales, es difícil, incluso imposible, deducir directamente una información pertinente sobre las distancias calculadas en las fronteras. En efecto, un valor de distancia puede corresponder a situaciones muy diferentes. Las figuras 3A a 3D ilustran de manera simple este principio. En estas figuras, se representa la frontera

$F_1$ , que separa un conjunto de vectores de picos  $\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$  de aprendizaje asociados a un primer microorganismo,

representados por unos círculos, de los otros vectores de picos  $\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$  de aprendizaje asociados a los otros microorganismos, representados por unos triángulos. Unos vectores de picos medidos  $M$  de un microorganismo desconocido a identificar son representados por unos cuadrados.

10

En el caso ilustrado en la figura 3A, la distancia del vector medido  $M$  a la frontera  $F_1$  es positiva. Sin embargo, el vector  $M$  está tan lejos del conjunto de vector de aprendizaje del primer microorganismo que no se puede deducir con certeza que el microorganismo desconocido es realmente el primer microorganismo. En el caso ilustrado en la figura 3B, el vector medido  $M$  está ahora próximo del conjunto de vector de aprendizaje pero también muy próximo de los otros conjuntos de vector de aprendizaje. En este caso, es también difícil deducir que el microorganismo desconocido es el primer microorganismo. En el caso ilustrado en la figura 3C, el vector medido  $M$  está lejos de la frontera  $F_1$  y está próximo del conjunto de vectores de aprendizaje, estando al mismo tiempo en el borde de este conjunto. Aunque este caso es más favorable que los casos anteriores, existe todavía una incertidumbre sobre la pertenencia del microorganismo a identificar. En particular, es necesario un estudio de la precisión de la medición. Finalmente, el caso ilustrado en la figura 3D es el caso típico y raro en el que el vector medido está al mismo tiempo lejos de la frontera y localizado entre el conjunto de los vectores de aprendizaje. La distancia medida es entonces un valor característico del primer microorganismo, en la que se puede tener confianza.

15

20

25 Como es posible constatar, las distancias calculadas son sólo parcialmente pertinentes. Por ejemplo, en un primer caso, una distancia igual a 0,4 es extremadamente pertinente, mientras que en otro caso es imposible deducir nada de ella. Por lo tanto, es necesario proceder a un análisis de estas distancias para deducir el tipo de microorganismo desconocido, así como el grado de confianza que se puede otorgar a esta identificación. Esta etapa de análisis suplementaria se realiza clásicamente por un operario, biólogo o médico, que determina con la ayuda de sus conocimientos qué conclusión es posible obtener de las distancias calculadas por la herramienta de clasificación.

30

Se ha descrito una clasificación vectorial, de tipo SVM, que calcula una distancia algébrica entre dos objetos de un espacio vectorial, a saber un vector que corresponde al microorganismo a identificar, y un hiperplano que corresponde a una frontera que divide el espacio, en el ejemplo ilustrado  $\mathbb{R}^2$ , en dos subespacios. El tipo de problema expuesto en relación con este tipo de clasificación se encuentra también en los otros tipos de clasificación a partir del momento en el que estas producen un valor, o resultado, que representa una distancia con objetos de referencia, que las clasificaciones sean de tipo SVM o no, o más generalmente de tipo vectorial o no, como por ejemplo las clasificaciones bayesianas, las clasificaciones lineales, las clasificaciones a base de redes neuronales, las clasificaciones a base de distancia tolerante, etc.

35

40

En una cierta medida, se puede argumentar que no existe todavía una herramienta fiable de identificación de microorganismos con la ayuda de una espectrometría de masas y de herramienta de clasificación que calcula unos valores de distancias.

45 Descripción de la invención

El objetivo de la presente invención es resolver el problema antes mencionado proponiendo un algoritmo de identificación de microorganismos a partir de mediciones de espectros de masa y de herramienta de clasificación, que permitan identificar de manera más fiable un microorganismo.

50

Para este propósito, la invención tiene por objeto un procedimiento de identificación por espectrometría de masas de un microorganismo entre un grupo predeterminado de microorganismos de referencia, siendo cada microorganismo de referencia representado por un conjunto de datos de referencia, comprendiendo el procedimiento:

55 \* la determinación de un conjunto de datos representativos del microorganismo a identificar en función de una medición de espectrometría de masas de dicho microorganismo; y

\* para cada microorganismo de referencia, el cálculo de una distancia entre dicho conjunto determinado de datos y el conjunto de datos de referencia del microorganismo de referencia.

60

Según la invención, el procedimiento comprende el cálculo de una probabilidad que el microorganismo a identificar

ya sea el microorganismo de referencia, según la relación:

$$f(m) = \frac{pN(m|\mu, \sigma)}{pN(m|\mu, \sigma) + (1-p)N(m|\bar{\mu}, \bar{\sigma})}$$

5 en la que:

\*  $m$  es la distancia calculada para el microorganismo de referencia;

\*  $f(m)$  es la probabilidad calculada para dicha distancia  $m$ ;

10 \*  $N(m|\mu, \sigma)$  es el valor, para la distancia  $m$ , de una variable aleatoria que modeliza la distancia entre un conjunto de datos asociado a un microorganismo a identificar y el conjunto de datos de referencia del microorganismo de referencia, cuando el microorganismo a identificar es el microorganismo de referencia;

15 \*  $N(m|\bar{\mu}, \bar{\sigma})$ , es el valor, para la distancia  $m$ , de una variable aleatoria que modeliza la distancia entre un conjunto de datos asociado a un microorganismo a identificar y el conjunto de datos de referencia del microorganismo de referencia, cuando el microorganismo a identificar no es el microorganismo de referencia; y

\*  $p$  es un escalar predeterminado comprendido entre 0 y 1.

20 Por "conjunto de datos de referencia" se entienden los datos que caracterizan un microorganismo de referencia en el sentido de una herramienta de clasificación. Por ejemplo, para una clasificación de tipo SVM-UTC, los datos de referencia asociado a un microorganismo de referencia corresponden a un hiperplano de un espacio vectorial que divide este espacio en dos sub-espacios.

25 Por "conjunto de datos representativos", se entienden los datos ilustrados para caracterizar un microorganismo a identificar en el sentido de la herramienta de clasificación elegida. Por ejemplo, por la clasificación de tipo SVM-UTC, estos datos son una lista de picos detectados que forma un vector del espacio vectorial.

30 En otras palabras, la invención se aplica a cualquier tipo de clasificación. Como es conocido en sí, una clasificación produce unas distancias que son unas cantidades objetivas que miden unas separaciones con unos elementos de referencia. Según la invención, estas distancias son transformadas en probabilidades normalizadas entre 0 y 1, según una ley de tipo "sigmoïde". Resulta que estas probabilidades son ellas mismas unas cantidades objetivas comparables entre sí, y por lo tanto medidas reales de la "similitud" de un microorganismo desconocido con un microorganismo previamente identificado.

35 Según un modo de realización, las variables aleatorias  $N(m|\mu, \sigma)$  y  $N(m|\bar{\mu}, \bar{\sigma})$  son unas variables aleatorias gáusicas, medios respectivamente iguales a  $\mu$  y  $\bar{\mu}$ , y distancias estándares respectivamente iguales a  $\sigma$  y  $\bar{\sigma}$ . Más particularmente, la probabilidad calculada según la relación:

40

$$f(m) = \frac{1}{1 + \frac{1-p}{p} \exp\left(\ln\left(\frac{\sigma}{\bar{\sigma}}\right) - \frac{1}{2(\sigma\bar{\sigma})^2} [(\sigma - \bar{\sigma})m - (\bar{\mu}\sigma - \mu\bar{\sigma})][(\sigma + \bar{\sigma})m - (\bar{\mu}\sigma + \mu\bar{\sigma})]\right)}$$

45 De manera ventajosa, si la relación lógica " $\mu > \bar{\mu}$ "  $\oplus$  " $\sigma > \bar{\sigma}$ " se verifica, en la que  $\oplus$  representa la función "o exclusivo", la probabilidad está fijada a 1 para cualquier distancia  $m$  superior a  $\frac{\bar{\mu}\sigma^2 - \mu\bar{\sigma}^2}{\sigma^2 - \bar{\sigma}^2}$ .

De manera ventajosa, si la relación lógica " $\mu > \bar{\mu}$ "  $\oplus$  " $\sigma > \bar{\sigma}$ " no se verifica, en la que  $\oplus$  representa la función "o exclusivo", la probabilidad se fija a 0 para cualquier distancia  $m$  inferior a  $\frac{\bar{\mu}\sigma^2 - \mu\bar{\sigma}^2}{\sigma^2 - \bar{\sigma}^2}$ .

Según un modo de realización, el escalar  $p$  es idéntico para todos los microorganismos de referencia. Más particularmente,  $p$  es igual a  $\frac{1}{N}$ , en la que  $N$  es el tamaño del conjunto de microorganismos de referencia. En una variante, el escalar  $p$  es igual a 0,5.

50

Según un modo de realización de la invención, la determinación del espectro de masas y el cálculo de la distancia entre el espectro adquirido y cada microorganismo de referencia utiliza un algoritmo de clasificación vectorial.

Más particularmente,

- 5
- \* la determinación del espectro de masas del microorganismo de masa comprende:
    - la adquisición de al menos un espectro de masas de dicho microorganismo;
    - 10 - la detección de picos en el al menos un espectro de masas adquirido y la conversión de los picos detectados en un vector de un espacio vectorial predeterminado,
  - \* y el cálculo de distancia entre dicho microorganismo y cada microorganismo de referencia comprende el cálculo de una distancia algébrica entre el vector determinado y una frontera que divide el espacio vectorial entre un primer sub-espacio característico del microorganismo de referencia y un segundo sub-espacio característico de los otros microorganismos de referencia.
- 15

En particular, la frontera de un microorganismo de referencia se calcula por medio de un algoritmo de tipo "máquina de soporte vectorial" y un conjunto de vectores correspondientes a los microorganismos de referencia.

- 20
- En una variante, la determinación del espectro de masas y el cálculo de la distancia entre el espectro adquirido y cada microorganismo de referencia utilizan un algoritmo de distancia tolerante.

- 25
- Según un modo de realización, el vector se calcula clasificando como máximo un pico en cada intervalo de una subdivisión predeterminada de la gama de las relaciones masa sobre carga del espectro de masas.

Breve descripción de las figuras

- 30
- La invención se entenderá mejor a la lectura de la descripción siguiente, dada únicamente a título de ejemplo, y realizada en relación con los dibujos anexos, en los que unas referencias idénticas designan unos elementos idénticos o análogos, y en los que:

- 35
- \* las figuras 1A, 1B y 1C son unos ejemplos de espectro de masas de dos picos, característicos respectivamente de 3 microorganismos;

- \* la figura 2 es un esquema que ilustra la obtención de las tres fronteras para los organismos de las figuras 1A, 1B, 1C por medio de un algoritmo de tipo SVM "uno contra todos";

- 40
- \* las figuras 3A a 3D son unos esquemas que ilustran cuatro distancias de un individuo ensayado en una frontera de tipo SVM;

- \* la figura 4 es un organigrama de un procedimiento según la invención;

- 45
- \* la figura 5 es un esquema que ilustra la determinación de dos variables aleatorias a partir de distancias de microorganismos a una frontera de tipo SVM;

- \* la figura 6 es un trazado que ilustra dos distribuciones gaussianas y una función de probabilidad obtenidas a partir de estas;

- 50
- \* las figuras 7 y 8 son unos trazados que ilustran la ausencia de carácter monótona que puede producirse en funciones de probabilidad según la invención;

- \* la figura 9 es un trazado que ilustra el resultado del forzamiento de una función de probabilidad según la invención en ser monótona;

- 55
- \* las figuras 10 y 11 son unas figuras que ilustran un caso de identificación difícil con la ayuda de la herramienta de clasificación y resuelto por la invención; y

- 60
- \* la figura 12 es un esquema que ilustra unas distancias tolerantes de un espectro a identificar con súper-espectros según el estado de la técnica.

Descripción detallada de la invención

- 65
- Se describirá ahora en relación con el organigrama de la figura 4, un procedimiento según la invención a base de clasificación SVM-UTC.

- El procedimiento empieza por una etapa 10 de adquisición de un conjunto de espectros de masa de aprendizaje y de un conjunto de espectros de masa de calibración de un nuevo microorganismo identificado a integrar en una base de conocimientos, por ejemplo mediante una espectrometría de masas de tipo MALDI-TOF (acrónimo de “*Matrix-assisted laser desorption/ionization time of flight*”). La espectrometría de masas MALDI-TOF es bien conocida en sí misma y no se describirá más en detalle a continuación. Se podrá hacer referencia, por ejemplo, al documento de Jackson O. Lay, «Maldi-tof spectrometry of bacteria», Mass Spectrometry Reviews, 2001, 20, 172-194. Los espectros adquiridos son después pretratados, a fin en particular de quitar el ruido de estos y quitar su línea de base, de una manera conocida en sí.
- Una identificación de los picos presentes en los espectros adquiridos se realiza entonces en 12, por ejemplo mediante un algoritmo de detección de picos basado en la detección de máximas locales. Se produce así una lista de los picos para cada espectro adquirido, que comprende la localización y la intensidad de los picos del espectro.
- De manera ventajosa, los picos se identifican en la gama de Thomson  $[m_{\min}; m_{\max}]$  predeterminada, preferentemente la gama  $[m_{\min}; m_{\max}] = [3000; 17000]$  Thomson. En efecto, se ha observado que las informaciones suficientes para la identificación de los microorganismos se agrupan en esta gama de relación en masa sobre carga, y que no es necesario por lo tanto tener en cuenta una gama más amplia.
- El procedimiento se prosigue, en 14, por una etapa de cuantificación, o “binning”. Para hacer esto, la gama  $[m_{\min}; m_{\max}]$  se subdivide en intervalos de anchura, por ejemplo constante, y para cada intervalo que comprende varios picos, se conserva un solo pico, ventajosamente el pico que presenta la mayor intensidad. Se produce así un vector para cada espectro medido. Cada componente del vector corresponde a un intervalo de la cuantificación y tiene por valor la intensidad del pico conservado para este intervalo, significando el valor “0” que no se ha detectado ningún pico en este intervalo.
- En una variante, los vectores son “binarizados” poniendo el valor de un componente del vector a “1” cuando un pico está presente en el intervalo correspondiente, y a “0” cuando no está presente ningún pico en este intervalo. Esto tiene por efecto hacer más firme la calibración de los algoritmos de clasificación realizada ulteriormente. Los inventores han notado en efecto que la información pertinente, en particular para la identificación de una batería, está contenida esencialmente en la ausencia y/o la presencia de picos, y que la información de intensidad es menos pertinente. Además, se observa que la intensidad es un tamaño muy variable de un espectro a otro y/o de un espectrómetro a otro. Debido a esta variable, es difícil tomar en cuenta los valores brutos de intensidad en las herramientas de clasificación.
- Paralelamente, los vectores de picos de espectros de aprendizaje, a continuación “vectores de aprendizaje”, y los vectores de picos de los espectros de calibración, a continuación “vectores de calibración” se memorizan en la base de conocimiento. Para cada uno de los microorganismos registrados en la base de conocimiento, existe por lo tanto un primer conjunto  $\{V_{i,j}^a\}$  de vectores de aprendizaje  $V_{i,j}^a$ , y un segundo conjunto  $\{V_{i,j}^c\}$  de vectores de calibración  $V_{i,j}^c$ .
- En una etapa 16 siguiente, una frontera  $F_j$  se calcula para cada microorganismo  $j$  de la base de datos entre el conjunto de los vectores de aprendizaje  $\{V_{i,j}^a\}$  del microorganismo y el conjunto  $\{V_{i,k \neq j}^a\}$  de los vectores de aprendizaje  $V_{i,k \neq j}^a$  de los otros microorganismos. Esta frontera  $F_j$  se calcula mediante un algoritmo de clasificación SVM-UTC. Este algoritmo es clásico y para más detalles se podrá hacer referencia, por ejemplo al documento de R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, y C.-J. Lin. «LIBLINEAR: A Library for Large Linear Classification», Journal of Machine Learning Research 9(2008), 1871-1874. Un programa que codifica este algoritmo está por ejemplo accesible en la dirección <http://www.csie.ntu.edu.tw/~cjlin/liblinear>.
- El procedimiento consiste después en determinar una función  $f_j$  que transforma la distancia algébrica a una frontera  $F_j$  en un grado de similitud con el microorganismo  $j$  asociado a esta frontera  $F_j$ . El principio subyacente a esta transformación es considerar la distancia de un microorganismo  $j$  a su frontera  $F_j$  asociada como siendo una primera variable aleatoria  $N_j(m|\mu_j, \sigma_j)$  y la distancia de los otros microorganismos a esta frontera como siendo una segunda variable aleatoria  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ , después en determinar una función  $f_j$  monótona, que varía de 0 a 1, que separa las dos variables aleatorias  $N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ .
- El procedimiento se prosigue por lo tanto por la identificación, en 18, de estas dos variables aleatorias,  $N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ . Para cada microorganismo  $j$  de la base de conocimientos, en función de los vectores de calibración

$\{V_{i,j}^c\}$  y  $\{V_{i,k \neq j}^c\}$ . Utilizando unos vectores de calibración diferentes de los vectores de aprendizaje utilizados para calcular las fronteras  $F_j$ , se evita un fenómeno de sobre-aprendizaje que puede ser perjudicial para la firmeza y la precisión de la identificación. Por supuesto, es también posible utilizar los vectores de aprendizaje para calcular las funciones  $f_j$ .

5 Volviendo al ejemplo de las figuras 1 y 2, en el que unos microorganismos son identificables por la localización de dos picos del espectro de masas, el cálculo de estas variables aleatorias se ilustra en la figura 5, en la que se representan unos vectores de calibración  $V_{i,j}^c$  de un microorganismo  $j$ , separado por una frontera  $F_j$  de los vectores de calibración  $V_{i,k \neq j}^c$  de los otros microorganismos de la base de conocimiento.

10 En primer lugar, se calcula la distancia  $m_{i,j}$  entre cada vector de calibración  $V_{i,j}^c$  del microorganismo  $j$  y la frontera  $F_j$ , así como la distancia  $\bar{m}_{i,j}$  entre cada vector de calibración  $V_{i,k \neq j}^c$  de los otros microorganismos de la base de conocimiento a dicha frontera  $F_j$ .

15 Después, la variable aleatoria  $N_j(m|\mu_j, \sigma_j)$  que modeliza la distancia entre un microorganismo  $j$  y su frontera  $F_j$  se identifica a partir del conjunto de las distancias  $m_{i,j}$ , y la variable aleatoria  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$  que modeliza la distancia entre los otros microorganismos y esta frontera  $F_j$  se identifica a partir del conjunto de las distancias  $\bar{m}_{i,j}$ .

20 De manera ventajosa, el modelo de variable aleatoria seleccionado es una distribución gaussiana. Las variables aleatorias  $N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$  se escriben por lo tanto según las relaciones:

$$N_j(m|\mu_j, \sigma_j) = \frac{1}{\sigma_j \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \left(\frac{m - \mu_j}{\sigma_j}\right)^2\right) \quad (1)$$

$$N_j(m|\bar{\mu}_j, \bar{\sigma}_j) = \frac{1}{\bar{\sigma}_j \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \left(\frac{m - \bar{\mu}_j}{\bar{\sigma}_j}\right)^2\right) \quad (2)$$

$$25 \quad \mu_j = \frac{1}{M_j^c} \sum_{i=1}^{M_j^c} m_{i,j} \quad (3)$$

$$\sigma_j = \sqrt{\frac{1}{M_j^c - 1} \sum_{i=1}^{M_j^c} (m_{i,j} - \mu_j)^2} \quad (4)$$

$$30 \quad \bar{\mu}_j = \frac{1}{M_j^{cneg}} \sum_{i=1}^{M_j^{cneg}} \bar{m}_{i,j} \quad (5)$$

$$\bar{\sigma}_j = \sqrt{\frac{1}{M_j^{cneg} - 1} \sum_{i=1}^{M_j^{cneg}} (\bar{m}_{i,j} - \bar{\mu}_j)^2} \quad (6)$$

35 en las que  $m$  es la distancia a la frontera  $F_j$ ,  $M_j^c$  es el número de vectores de calibración  $V_{i,j}^c$  del microorganismo  $j$ , y  $M_j^{cneg}$  es el número total de vectores de calibración  $V_{i,k \neq j}^c$  de los otros microorganismos de la base de conocimiento.

El procedimiento se prosigue mediante el cálculo, en 20, de una función  $f_j$  de probabilidad para cada microorganismo  $j$  que separa las distribuciones,  $N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ , tal como se ilustra en la figura 6.

40

Más particularmente, la función  $f_j$  satisface la relación:

$$f_j(m) = \frac{p_j \times N_j(m|\mu_j, \sigma_j)}{p_j \times N_j(m|\mu_j, \sigma_j) + (1-p_j) \times N_j(m|\bar{\mu}_j, \bar{\sigma}_j)} \quad (7)$$

5 En la que  $p_j$  es un parámetro de ajuste predeterminado relativo a una información a priori sobre la prevalencia del microorganismo  $j$ .

El parámetro  $p_j$  es por ejemplo igual a 0,5 si el evento "la distancia  $m$  corresponde al microorganismo  $j$ " tiene el mismo peso que el evento "la distancia  $m$  corresponde a un microorganismo del microorganismo  $j$ ". En esta hipótesis, se considera que si  $m$  se encuentra a equidistancia de las 2 distribuciones, el espectro tiene una

10 probabilidad  $\frac{1}{2}$  de ser el microorganismo de referencia  $j$ , y por lo tanto una probabilidad de  $\frac{1}{2(N-1)}$  de ser cualquiera de los  $N-1$  diferentes de los microorganismos de referencia, en la que  $N$  es el número de microorganismos registrados en la base de datos.

15 En una variante, el parámetro  $p_j$  es igual a  $\frac{1}{N}$ , en la que  $N$  es el número de microorganismos registrado en la base de conocimientos. En esta hipótesis, se considera que si  $m$  se encuentra a equidistancia de las 2 distribuciones, el espectro tiene la probabilidad de estar en cualquiera de los microorganismos de referencia.

Como se ilustra en la figura 6, la función de probabilidad  $f_j$  es una función monótona al menos sobre un intervalo de

20 distancias que comprende lo esencial de las distribuciones,  $N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ , y que crece de 0 a 1 tomando forma de un sigmoideo. En particular, para los valores importantes de,  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ , esta función es igual a 0 mientras que para los valores importantes de,  $N_j(m|\mu_j, \sigma_j)$  esta función es igual a 1.

La función  $f_j$  representa en efecto un grado de probabilidad, de que un microorganismo que tiene una distancia  $m$  en la frontera  $F_j$  sea el microorganismo  $j$ . Más particularmente, se muestra que la función  $f_j$  corresponde a la probabilidad  $p(s|m)$  que un microorganismo ensayado sea el microorganismo  $j$  cuando se observa la distancia  $m$ , es decir que la función  $f_j$  según la relación 7 corresponde a la probabilidad  $p(s|m)$  según las relaciones:

$$p(s|m) = \frac{p(s)p(m|s)}{p(m)} \quad (8)$$

$$30 \quad p(m) = p(s)p(m|s) + p(\bar{s})p(m|\bar{s}) \quad (9)$$

en las que  $p(s)$  es una distribución a priori sobre la probabilidad de obtener el microorganismo  $j$ , es decir su prevalencia,  $p(m|\bar{s})$  es la verosimilitud de la distancia  $m$  para el microorganismo  $j$ , y  $p(m|\bar{s})$  es la verosimilitud los

35 otros microorganismos. Según la invención, se tiene por lo tanto  $p(s) = p_j$ ,  $p(m|s) = N_j(m|\mu_j, \sigma_j)$  y  $p(m|\bar{s}) = N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ .

Si el caso de las distribuciones expuestas a la figura 6 es un caso estándar, puede a veces pasar que la función  $f_j$  no sea monótona sobre una parte extrema de una distribución, como se representa en las figuras 7 y 8.

40 En efecto, el modelo de distribución,  $N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$  elegido puede ser impreciso, o el número de muestras utilizadas para generar los vectores de calibración  $\{V_{i,j}^c\}$  y  $\{V_{i,k7j}^c\}$  ser limitado, y por lo tanto inducir una incertidumbre sobre los parámetros de distribuciones, o también de los valores aberrantes de distancia pueden producirse, por ejemplo debido a incertidumbre de medición.

45 En particular, la función de transformación en probabilidades  $f_j$  calculada a partir de leyes normales tales como se definen en las relaciones (1) a (6) puede no ser monótona. En particular, las dos distribuciones normales



$N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ , tienen unas distancias estándares  $\sigma_j$  y  $\bar{\sigma}_j$  lo más frecuentemente muy diferentes, siendo la distribución  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$  asociada a los no representantes del microorganismo  $j$  generalmente mucho más estrecha que la distribución  $N_j(m|\mu_j, \sigma_j)$  de los representantes del microorganismo  $j$ . Esto se traduce frecuentemente por un carácter no monótono de la función  $f_j$  sobre el conjunto de las distancias, en particular por una subida de la función  $f_j$  para las distancias en la "izquierda" de la distribución  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ , como se ilustra en las figuras 7 y 8.

Por ejemplo, en la figura 7, para una distancia  $m_{375}$  entre un microorganismo a identificar y la frontera  $F_{375}$  asociada al microorganismo nº 375 de la base de conocimiento igual a -2, sin medición particular, la función  $f_{375}$  del microorganismo nº 375 es igual a 1. Esto significa que el microorganismo a identificar es el microorganismo nº 375 con una probabilidad próxima del 100%, a pesar de que parece, por el contrario, que la probabilidad real que el microorganismo medido sea el microorganismo nº 375 sea muy baja. Además, existen otras funciones  $f_j$  para las cuales el valor será positivo para la distancia  $m$  del microorganismo a identificar en la frontera  $F_j$  asociada al microorganismo  $j$ , y por lo tanto una probabilidad asociada muy superior a 0. Por ejemplo, en la figura 8, la probabilidad  $f_{375}(m_{375})$  del microorganismo a identificar, asociada a la distancia  $m_{375}$  de este último a la frontera  $F_{1515}$  de un microorganismo nº 1515 de la base de datos, es igual a 0,9, es decir una probabilidad inferior a la asociada al microorganismo nº 375, mientras que parece que el microorganismo a identificar es efectivamente el microorganismo nº 1515. Conviene por lo tanto que la confianza en esta identificación del microorganismo nº 1515 sea superior a la asociada al microorganismo nº 375.

A pesar de que en la realidad, es poco probable obtener unas distancias medidas en partes extremas de las funciones  $f_j$  en las que podría existir tal comportamiento, la función  $f_j$  está ventajosamente determinada a fin de garantizar su carácter monótono.

Cabe señalar para ello que las distribuciones gaussianas se seleccionan ventajosamente en razón a la simplicidad de su cálculo, y que el comportamiento no monótono que puede, en algunos casos, resultar de esta elección, no es crítico. En efecto, la utilización de gaussiano presenta la ventaja de producir unas funciones  $f_j$  cuyo comportamiento es analíticamente predecible. Por lo tanto, es posible de modificar las funciones  $f_j$  según la relación (7) a fin de garantizar en la salida un comportamiento monótono sobre el conjunto de las distancias.

Más particularmente, se muestra que, para una función  $f_j$  a base de distribuciones gaussianas, puede sólo existir un único mínimo local de valor según la relación:

$$e = \frac{\bar{\mu}_j \sigma_j^2 - \mu_j \bar{\sigma}_j^2}{\sigma_j^2 - \bar{\sigma}_j^2} \quad (10)$$

De manera ventajosa, el procedimiento se prosigue por lo tanto por una etapa 22 en la que las funciones  $f_j$  se vuelven monótonas. Más particularmente, se plantea:

$$f_j(m) = 0 \text{ para todo } m \leq e, \text{ si } \mu > \bar{\mu}, \text{ o de manera equivalente } \sigma > \bar{\sigma} \quad (11)$$

$$f_j(m) = 1 \text{ para todo } m \geq e, \text{ si la relación lógica } "\mu > \bar{\mu}" \oplus "\sigma > \bar{\sigma}" \text{ es verdadera} \quad (12)$$

en la que  $\oplus$  es el símbolo de la función lógica "o exclusiva".

Una aplicación de la etapa 22 está, por ejemplo, ilustrada en la figura 9, en la que la función  $f_{375}$  ilustrada en la figura 7 está forzada al valor nulo para cualquier distancia inferior al valor  $e$ , de manera que la función  $f_{375}$  es ahora una función monótona de tipo sigmoideo sobre el conjunto de las distancias.

Al final de la etapa 22, la herramienta de clasificación según la invención está por lo tanto calibrada. Las fronteras  $F_j$  de la herramienta de clasificación SVM-UCT así como las funciones de probabilidad  $f_j$  son memorizadas en la base de conocimiento.

La base de conocimiento se incorpora en un sistema de identificación de microorganismos por espectrometría de masas que comprende un espectrómetro de masa, por ejemplo un espectrómetro de tipo MALDI-TOF, así como una unidad de tratamiento de informaciones, conectada al espectrómetro y apto para recibir y tratar los espectros de masa adquiridos para identificar unos microorganismos desconocidos.

Más particularmente, para la identificación de un microorganismo desconocido, el procedimiento comprende una etapa 24 de adquisición de uno o varios espectros de masa de este, una etapa 26 de pretratamiento de os espectros

adquiridos así como una etapa 28 de detección de los picos de los espectros y de determinación de un vector de picos  $V_m$ , tal como, por ejemplo, descritos anteriormente en relación con las etapas 10 a 14.

5 En una etapa 30 siguiente, la distancia  $m_j$  del vector  $V_m$  a cada una de las fronteras  $F_j$  se calcula, después cada una de las distancias  $m_j$  se transforma, en la etapa 32, en la probabilidad correspondiente  $f_j(m_j)$ .

En una primera variante, las probabilidades  $f_j(m_j)$  se calculan numéricamente según las fórmulas (1), (2) y (7), a saber según la relación de tipo:

$$f_j(m) = \frac{p_j \times \frac{1}{\sigma_j \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \left(\frac{m - \mu_j}{\sigma_j}\right)^2\right)}{p_j \times \frac{1}{\sigma_j \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \left(\frac{m - \mu_j}{\sigma_j}\right)^2\right) + (1 - p_j) \times \frac{1}{\bar{\sigma}_j \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \left(\frac{m - \bar{\mu}_j}{\bar{\sigma}_j}\right)^2\right)} \quad (13)$$

10 En esta variante, el numerador y el denominador se calculan independientemente el uno del otro utilizando unas funciones exponenciales numéricas. Sin embargo, cuando un microorganismo presenta una distancia  $m_j$  lejos de los medios  $\mu_j$  y  $\bar{\mu}_j$  de las dos distribuciones gaussianas, existe un riesgo de que los exponenciales numéricos resultantes sean aproximados a 0 y que resulte una indeterminación de tipo 0/0 para el valor  $f_j(m_j)$ .

Según una segunda variante, las probabilidades  $f_j(m_j)$  se calculan según la relación (14) siguiente:

$$f_j(m) = \frac{1}{1 + \frac{1 - p_j}{p_j} \exp\left(\ln\left(\frac{\sigma_j}{\bar{\sigma}_j}\right) - \frac{1}{2(\sigma_j \bar{\sigma}_j)} \left[ (\sigma_j - \sigma_j)m_j - (\mu_j \sigma_j - \mu_j \sigma_j) \right] \left[ (\sigma_j + \sigma_j)m_j - (\mu_j \sigma_j + \mu_j \sigma_j) \right] \right)}$$

20 La relación (13) es, desde el punto de vista matemático, estrictamente equivalente a la relación (14). Sin embargo, la relación (14) es más firme a las aproximaciones numéricas, y soporta incluso unas distancias de valor infinito sin producir indeterminación.

25 En el caso en el que las funciones  $f_j$  son forzadas monótonas, también se aplican las relaciones (11) y (12).

Las probabilidades  $f_j(m_j)$  se clasifican entonces por orden decreciente. Sin ninguna supera un cierto límite, se considera por ejemplo que no hay identificación. Si por el contrario una o varias superan un cierto límite, son por ejemplo mostrada, en 34, sobre una pantalla del sistema de identificación con una lista de microorganismos asociados, por ejemplo una lista de 3 microorganismos se muestra con su valor de  $f_j(m_j)$  correspondiente que constituye directamente el grado de confianza que se puede dar a la identificación de los microorganismos.

Las figuras 10 y 11 ilustran un caso de identificación particularmente difícil de un microorganismo por la herramienta de clasificación y resuelto por la invención. La figura 10 es un trazado que ilustra unas distribuciones gaussianas  $N_{325}(m|\mu_{325}, \sigma_{325})$  y  $N_{325}(m|\bar{\mu}_{325}, \bar{\sigma}_{325})$ , así como sus funciones  $f_{325}$  asociadas, que corresponden a un organismo nº 325 de la base de conocimiento, y la figura 11 es un trazado que ilustra unas distribuciones gaussianas  $N_{59}(m|\mu_{59}, \sigma_{59})$  y  $N_{59}(m|\bar{\mu}_{59}, \bar{\sigma}_{59})$ , así como sus funciones  $f_{59}$ , que corresponden a un organismo nº 59 de la base de conocimiento. Una distancia  $m_{59}$  de un microorganismo a identificar en la frontera  $F_{59}$  del microorganismo nº 59 es igual a -1,1, mientras que la distancia  $m_{325}$  del microorganismo a identificar en la frontera  $F_{325}$  del microorganismo nº 325 es igual a -0,9.

Conforme al estado de la técnica, que se basa en la comparación directa de las distancias, sin análisis suplementario, se concluye por lo tanto que el microorganismo a identificar es el microorganismo nº 325. Ahora bien, mirando las distribuciones, se observa que el microorganismo a identificar no tiene prácticamente ninguna posibilidad de ser el microorganismo nº 325, mientras que tiene muchas posibilidades de ser el microorganismo nº 59. Sólo un análisis suplementario, basado en el conocimiento de la persona encargada de la identificación, es capaz de concluir a partir de las distancias medidas que es en realidad el microorganismo nº 59 que se ha identificado.

50 La transformación de las distancias en función de probabilidad según la invención permite resolver este tipo de caso complejo. En efecto, la función  $f_{325}(m_{325})$  es nula, lo que significa que la probabilidad que el microorganismo a identificar sea el microorganismo nº 325 es nula, mientras que la función  $f_{59}(m_{59})$  es positiva, aquí igual a 0,52, lo que

significa que el microorganismo a identificar es el microorganismo nº 59 con una confianza del 52%.

De manera ventajosa, el procedimiento según la invención permite también juzgar si sí o no el microorganismo desconocido es uno de los microorganismos de referencia de la base de conocimiento. En efecto, los valores  $f_j(m)$  representan unas mediciones de la similitud de este microorganismo con los microorganismos de referencia. Así, si los valores  $f_j(m)$  son bajos, se puede juzgar que el microorganismo desconocido no es similar a ninguno de los microorganismos de referencia, y por lo tanto no está referenciado en la base de datos. Más particularmente, el procedimiento según la invención comprende una etapa de comparación de cada uno de los valores  $f_j(m)$  con un valor límite  $s_j$  predeterminado, por ejemplo único a todos los microorganismos de referencia o un valor particular para cada uno de ellos, y si los valores  $f_j(m)$  son todos inferiores a sus valores límites respectivos, entonces se determina que el microorganismo de referencia no corresponde a ninguno de los microorganismos de referencia de la base de conocimiento.

De manera ventajosa, se utiliza un único valor límite  $s_j$ , siendo este valor igual a 60%. Los inventores han constatado, en efecto, que este valor permite determinar con certeza que un microorganismo desconocido no pertenece a la base de conocimiento.

De manera ventajosa, el procedimiento reenvía también los valores  $f_j(m)$  próximos, por ejemplo los valores  $f_j(m)$  máximos que difieren el uno del otro en menos del 10%.

A pesar de que el procedimiento según la invención permite resolver un número importante de casos delicados que ponen en dificultad el clasificador en base al cual se determinan las funciones de probabilidad  $f_j$ , existe no obstante un número de casos en el que el procedimiento según la invención reenvía unos resultados similares para diferentes organismos de referencia, en particular cuando unos microorganismos de referencia son muy similares, por ejemplo en términos de fenotipo. Al ser los valores  $f_j(m)$  unas mediciones de similitud entre microorganismos, el procedimiento según la invención permite por lo tanto reenviar unos resultados de valor sustancialmente próximo para unos microorganismos de referencia intrínsecamente similares que el clasificador no ha podido distinguir con un margen mínimo.

Se ha descrito un modo de realización particular de la invención en el que se utilizan unas distribuciones gaussianas

para modelizar las variables aleatorias  $N_j(m|\mu_j, \sigma_j)$  y  $N_j(m|\bar{\mu}_j, \bar{\sigma}_j)$ .

Otros tipos de variables aleatorias son por supuesto posibles, siendo lo esencial obtener una modelización pertinente de las distribuciones. Por ejemplo, es posible seleccionar unas mezclas gaussianas, para tener en cuenta la presencia eventual de varias modalidades en las distribuciones. Las distribuciones gaussianas son ventajosas en la medida en la que su identificación es muy simple y que se ha constatado que esta modelización, incluso imperfecta, permitía la obtención de funciones de probabilidad  $f_j$  firmes.

Asimismo, se ha descrito un modo de realización en el que la herramienta de clasificación utilizado es un algoritmo de tipo SVM "uno contra todos".

Por supuesto, la invención no está limitada a este tipo de algoritmo, y se aplica a cualquier tipo de algoritmo de clasificación, en particular a los algoritmos de clasificación multi-clases realizados por combinación de clasificaciones binarias de tipo "uno contra todos" o "todos contra todos" por ejemplo, ya que el algoritmo de clasificación produce un valor, o resultado, que representa una distancia a unos objetos de referencia, que no es directamente un índice de confianza.

Como se conoce en sí, sea cual sea el algoritmo de clasificación considerado, existe siempre unos resultados o distancias calculadas con respecto a unos datos que representan cada uno de los microorganismos de referencia, en base a los cuales el procedimiento según la invención puede apoyarse.

En particular, la invención se aplica a unos algoritmos de clasificación que aplican un cálculo de similitud con respecto a unos elementos de referencia, como por ejemplo unos espectros medios, o unos "super-espectros" tales como se describen en el documento EP 1 253 622.

En el algoritmo descrito en el documento EP 1 253 622 B1, unos "espectros de referencia sintética" (REFs), denominados también "súper-espectros" se construyen y corresponden cada uno a una lista de picos reputados los más típicos de una especie dada. Para identificar un microorganismo con la ayuda de un espectro de masas, se calcula una "similitud" de este espectro medido a todos los súper-espectros memorizados en una base de conocimiento. Esta similitud puede, por ejemplo, ser una distancia que integra una tolerancia sobre las masas, como por ejemplo una distancia de Jaccard o de Hamming.

En particular, si  $SSp_1$  es un súper-espectro compuesto de una lista de masas  $M_i$  y de pesos asociados  $W_i$ , y  $Sp_1$  es una lista de masas  $M'_j$  asociadas a unos picos detectados en el espectro de masas del microorganismo a identificar,

entonces se calcula la similitud, o distancia tolerante, entre  $SSpl$  y  $SpI$ , como se conoce en sí misma, según la relación:

$$s(SSpl/SpI) = \frac{\sum W_i, \exists M'_j, \frac{|M'_j - M_i|}{M_i} < d}{\sum W_i} \quad (15)$$

5

en la que  $d$  es la tolerancia relativa aceptada entre dos masas.

10

Según el estado de la técnica, se retiene como identificación el organismo que corresponde al súper-espectro que tiene la similitud más alta, sirviendo este valor de similitud directamente de medición de la confianza que se puede obtener en el resultado. Ahora bien, según la elección de los picos que han servido para construir el súper-espectro, las similitudes medias para el conjunto de los espectros de una especie con su súper-espectro (REFs) no son obligatoriamente las mismas. Así, en el ejemplo ilustrado en la figura 12, a pesar de que el espectro  $SpI$  sea equidistante de los 3 súper-espectros  $A$ ,  $B$ ,  $C$ , se comprende que la probabilidad que sea efectivamente de la especie  $A$  es inferior a la que sea de la especie  $B$ , como aparece claramente a la vista de la distribución de los

15

espectros muestras  $\{Sp_A^c\}$ ,  $\{Sp_B^c\}$ ,  $\{Sp_C^c\}$  que han servido de base a la construcción de los súper-espectros. La invención permite normalizar las similitudes obtenidas y así levantar esta ambigüedad.

**REIVINDICACIONES**

1. Procedimiento de identificación por espectrometría de masas de un microorganismo entre un conjunto predeterminado de  $J$  microorganismos de referencia, siendo  $J$  un número entero positivo, siendo cada microorganismo de referencia representado por un conjunto de datos de referencia predeterminados por un algoritmo de clasificación, comprendiendo el procedimiento:

- la determinación de un conjunto de datos representativos del microorganismo a identificar en función de una medición de espectrometría de masas de dicho microorganismo; y
- para cada microorganismo de referencia  $j, j \in [1, J]$ , el cálculo de una distancia  $m_j$  entre dicho conjunto determinado de datos y el conjunto de datos de referencia del microorganismo de referencia  $j$ ,

caracterizado por que comprende el cálculo de una probabilidad  $f_j(m_j)$  que el microorganismo a identificar es el microorganismo de referencia  $j$ , según la relación:

$$f_j(m_j) = \frac{p_j \times N_j(m_j)}{p_j \times N_j(m_j) + (1 - p_j) \times \bar{N}_j(m_j)}$$

en la que:

- $m_j$  es la distancia calculada entre el conjunto de datos representativos del microorganismo a identificar y el conjunto de datos de referencia del microorganismo de referencia  $j$ ;
- $f_j(m_j)$  es la probabilidad calculada para dicha distancia  $m_j$ ;
- $N_j(m_j)$  es el valor, para la distancia  $m_j$ , de una variable aleatoria que modeliza la distancia entre un conjunto de datos asociados a un microorganismo a identificar y el conjunto de datos de referencia del microorganismo de referencia  $j$ , cuando el microorganismo a identificar es el microorganismo de referencia;
- $\bar{N}_j(m_j)$  es el valor, para la distancia  $m_j$ , de una variable aleatoria que modeliza la distancia entre un conjunto de datos asociados a un microorganismo a identificar y el conjunto de datos de referencia del microorganismo  $j$  de referencia, cuando el microorganismo a identificar no es el microorganismo de referencia  $j$ ;
- $p_j$  es un escalar predeterminado comprendido entre 0 y 1.

2. Procedimiento de identificación según la reivindicación 1, caracterizado por que las variables aleatorias  $N_j(m_j)$  y  $\bar{N}_j(m_j)$  son unas variables aleatorias gaussianas, de medios respectivamente iguales a  $\mu_j$  y  $\bar{\mu}_j$ , y de desviaciones estándares respectivamente iguales a  $\sigma_j$  y  $\bar{\sigma}_j$ .

3. Procedimiento de identificación según la reivindicación 2, caracterizado por que la probabilidad se calcula según la relación:

$$f_j(m) = \frac{1}{1 + \frac{1-p_j}{p_j} \exp\left(\ln\left(\frac{\sigma_j}{\bar{\sigma}_j}\right) - \frac{1}{2(\sigma_j \bar{\sigma}_j)^2} [(\sigma_j - \bar{\sigma}_j)m_j - (\bar{\mu}_j \sigma_j - \mu_j \bar{\sigma}_j)] [(\sigma_j + \bar{\sigma}_j)m_j - (\bar{\mu}_j \sigma_j + \mu_j \bar{\sigma}_j)]\right)}$$

4. Procedimiento de identificación según la reivindicación 2 o 3, caracterizado por que, si la relación lógica " $\mu_j > \bar{\mu}_j$ "  $\oplus$  " $\sigma_j > \bar{\sigma}_j$ " se verifica, en la que  $\oplus$  representa la función "o exclusivo", la probabilidad se fija a 1 para

$$\frac{\bar{\mu}_j \sigma_j^2 - \mu_j \bar{\sigma}_j^2}{\sigma_j^2 - \bar{\sigma}_j^2}$$

cualquier distancia  $m_j$  superior a

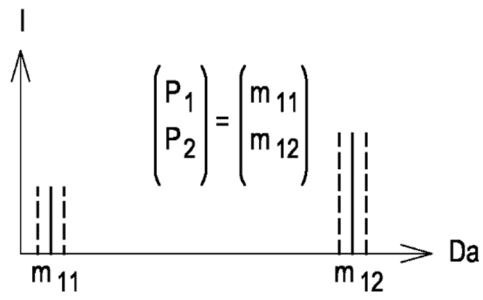
5. Procedimiento de identificación según la reivindicación 2 o 3, caracterizado por que si la relación lógica " $\mu_j > \bar{\mu}_j$ "  $\oplus$  " $\sigma_j > \bar{\sigma}_j$ " no está verificada, en la que  $\oplus$  representa la función "o exclusivo" la probabilidad se fija a

$$\frac{\bar{\mu}_j \sigma_j^2 - \mu_j \bar{\sigma}_j^2}{\sigma_j^2 - \bar{\sigma}_j^2}$$

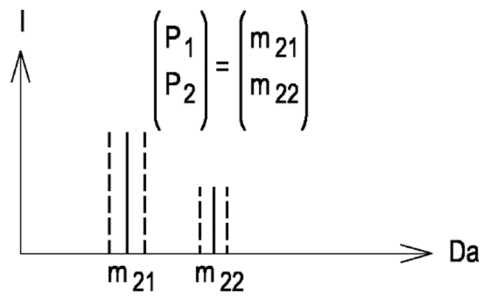
0 para cualquier distancia  $m_j$  inferior a

6. Procedimiento de identificación según una cualquiera de las reivindicaciones anteriores, caracterizado por que el escalar  $p_j$  es idéntico para todos los microorganismos de referencia.

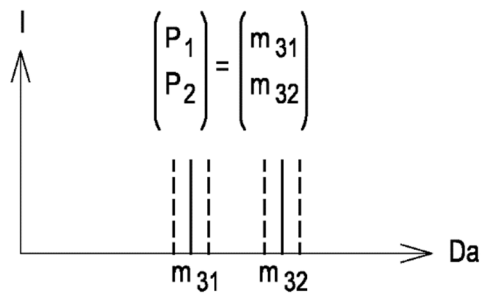
7. Procedimiento de identificación según la reivindicación 6, caracterizado por que el escalar  $p_j$  es igual a  $\frac{1}{N}$ , en la que  $N$  es el tamaño del conjunto de microorganismos de referencia.
- 5 8. Procedimiento de identificación según una cualquiera de las reivindicaciones 1-6, caracterizado por que el escalar  $p_j$  es igual a 0,5.
9. Procedimiento de identificación según una cualquiera de las reivindicaciones anteriores, caracterizado por que la determinación del espectro de masas y el cálculo de la distancia entre el espectro adquirido y cada microorganismo de referencia utilizan un algoritmo de clasificación vectorial.
- 10 10. Procedimiento de identificación según la reivindicación 9, caracterizado:
- por que la determinación del espectro de masas del microorganismo de masa comprende:
    - la adquisición de al menos un espectro de masas de dicho microorganismo;
    - la detección de picos en el al menos un espectro de masas adquirido y la conversión de los picos detectados en un vector de un espacio vectorial predeterminado,
    - y por que el cálculo de distancia entre dicho microorganismo y cada microorganismo de referencia comprende el cálculo de una distancia algébrica entre el vector determinado y una frontera que divide el espacio vectorial entre un primer sub-espacio característico del microorganismo de referencia y un segundo sub-espacio característico de los otros microorganismos de referencia.
- 15 20 25
11. Procedimiento de identificación según la reivindicación 9 o 10, caracterizado por que la frontera de un microorganismo de referencia se calcula mediante un algoritmo de tipo "soporte vector máquina" y de un conjunto de vectores que corresponden a los microorganismos de referencia.
- 30 12. Procedimiento de identificación según una de las reivindicaciones 9, 10 o 11, caracterizado por que el vector se calcula clasificando como máximo un pico en cada intervalo de una subdivisión predeterminada de la gama de las relaciones masa-sobrecarga del espectro de masas.
- 35 13. Procedimiento de identificación según una cualquiera de las reivindicaciones anteriores, caracterizado por que la determinación del espectro de masas y el cálculo de la distancia entre el espectro adquirido y cada microorganismo de referencia utilizan un algoritmo de distancia tolerante.
- 40 14. Procedimiento de identificación según una cualquiera de las reivindicaciones anteriores, caracterizado por que comprende:
  - la comparación de cada una de las probabilidades  $f_j(m_j)$  con un valor límite predeterminado; y
  - la determinación de que el microorganismo a identificar no corresponde a ninguno de los microorganismos de referencia si todas las probabilidades  $f_j(m_j)$  son inferiores al valor límite.
- 45 15. Procedimiento de identificación según la reivindicación 14, caracterizado por que el valor límite es igual al 60%.
16. Dispositivo de identificación de un microorganismo por espectrometría de masas, que comprende:
  - un espectrómetro de masa apto para producir unos espectros de masa de microorganismos a identificar;
  - una unidad de cálculo apta para identificar los microorganismos asociados a los espectros de masa producidos por el espectrómetro utilizando un procedimiento conforme a una cualquiera de las reivindicaciones 1-15.
- 50



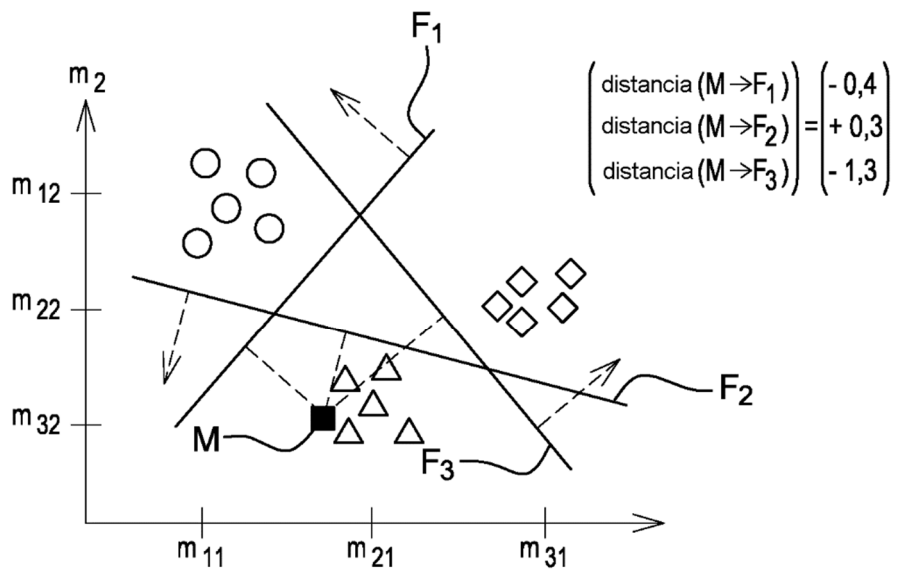
**Fig. 1A**



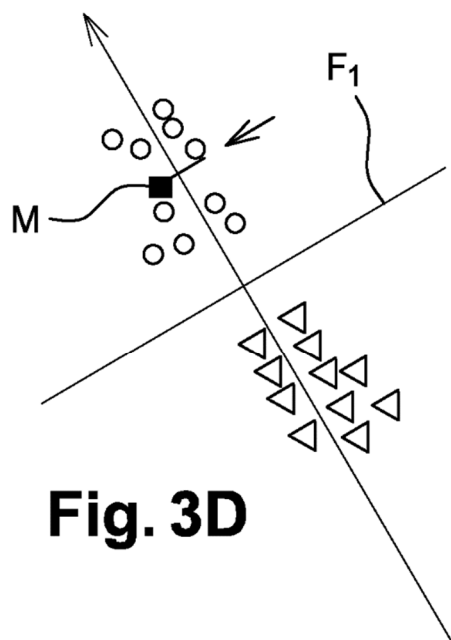
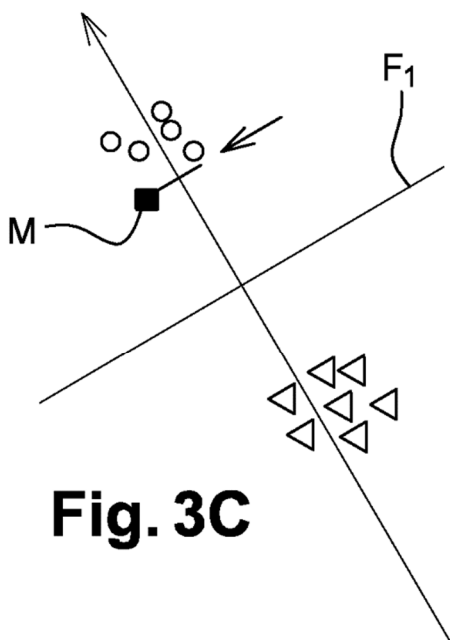
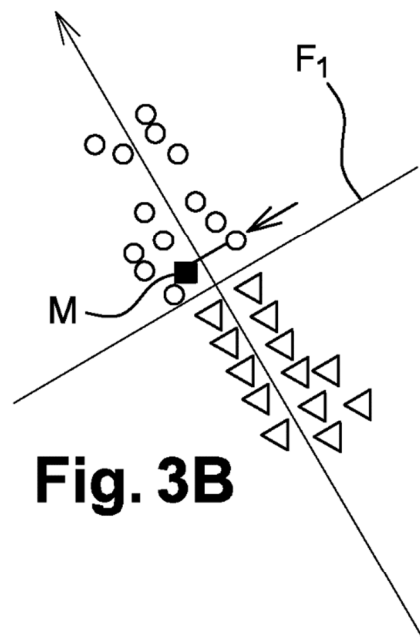
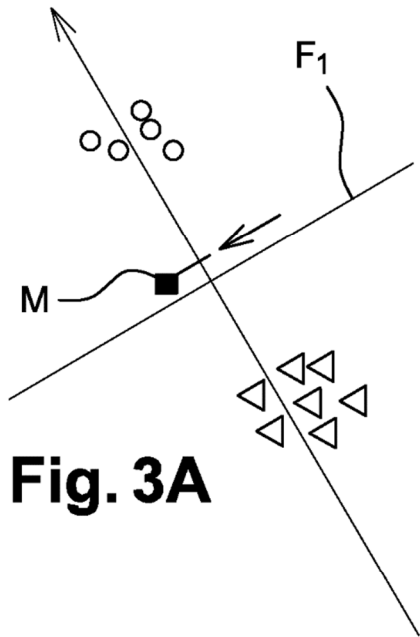
**Fig. 1B**



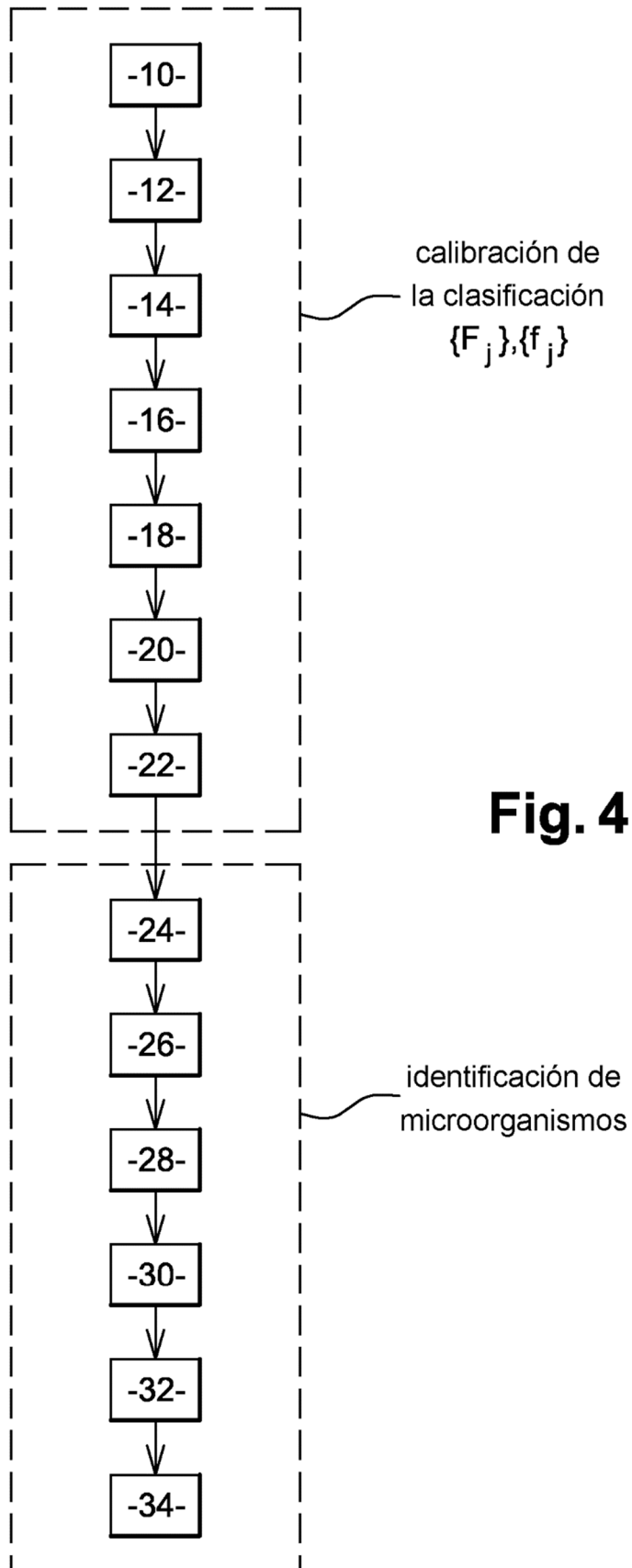
**Fig. 1C**



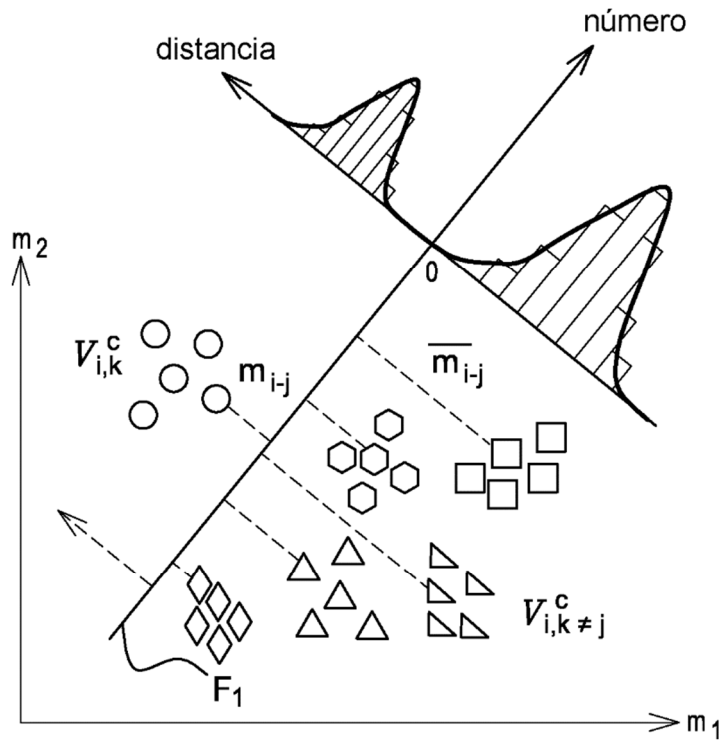
**Fig. 2**



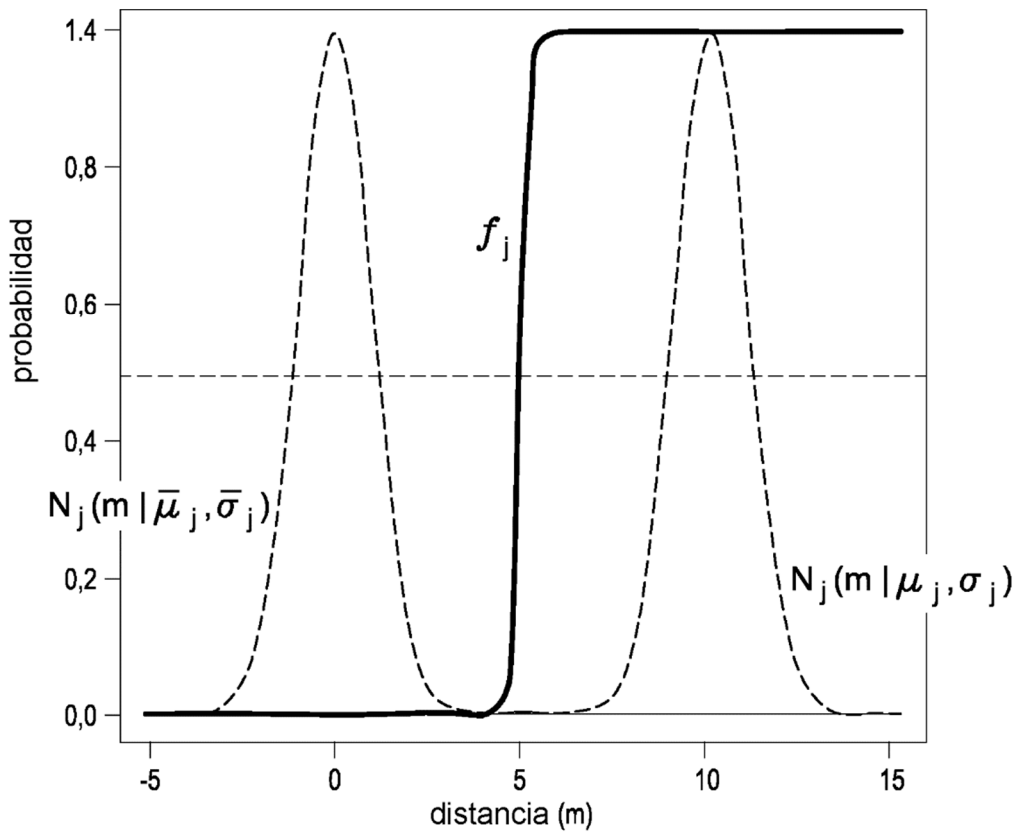




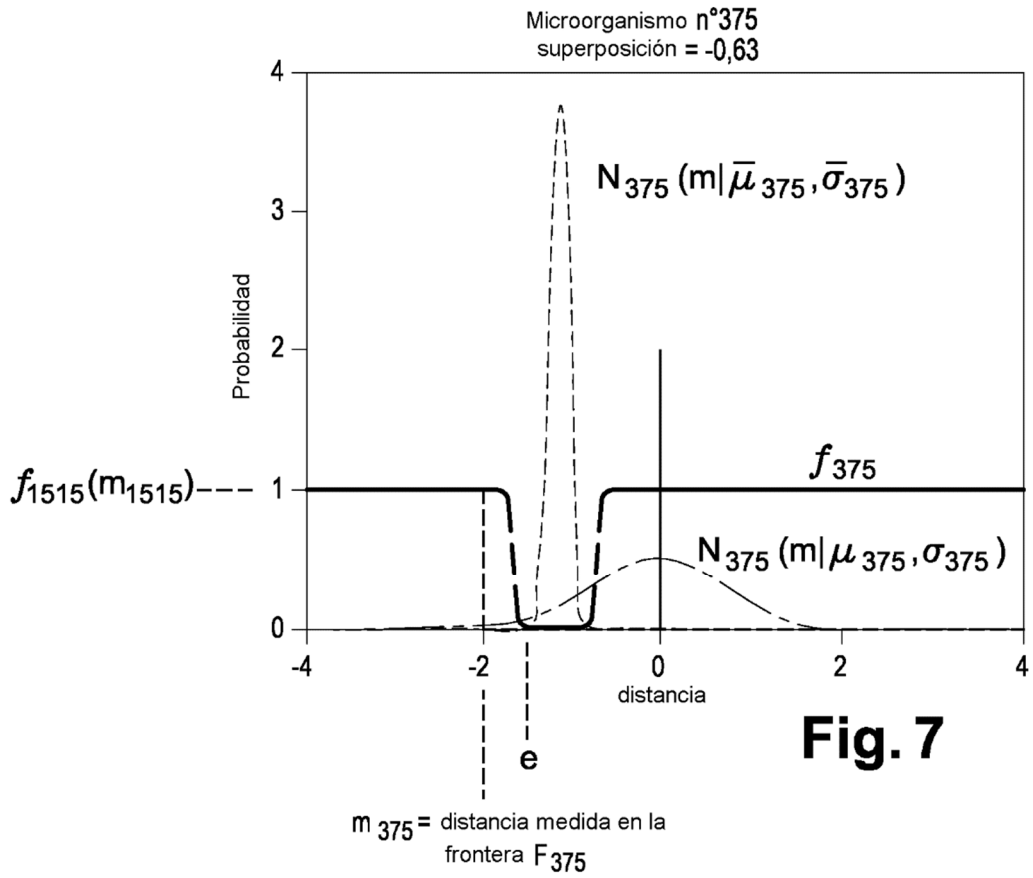
**Fig. 4**



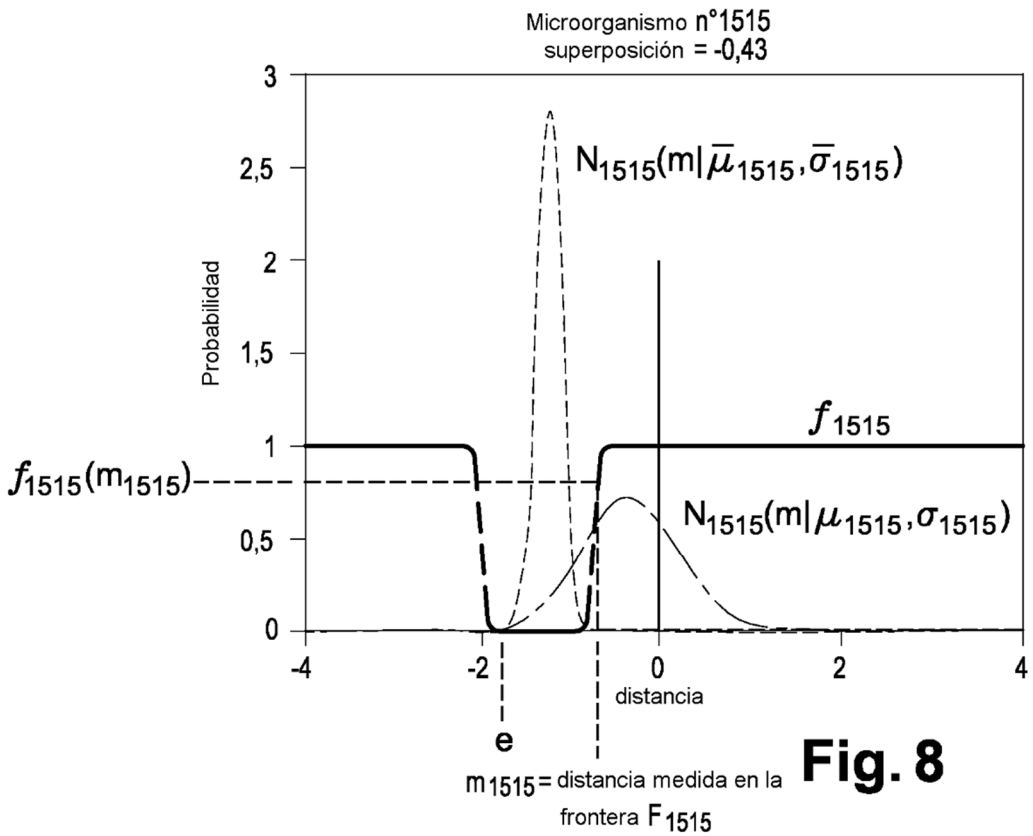
**Fig. 5**



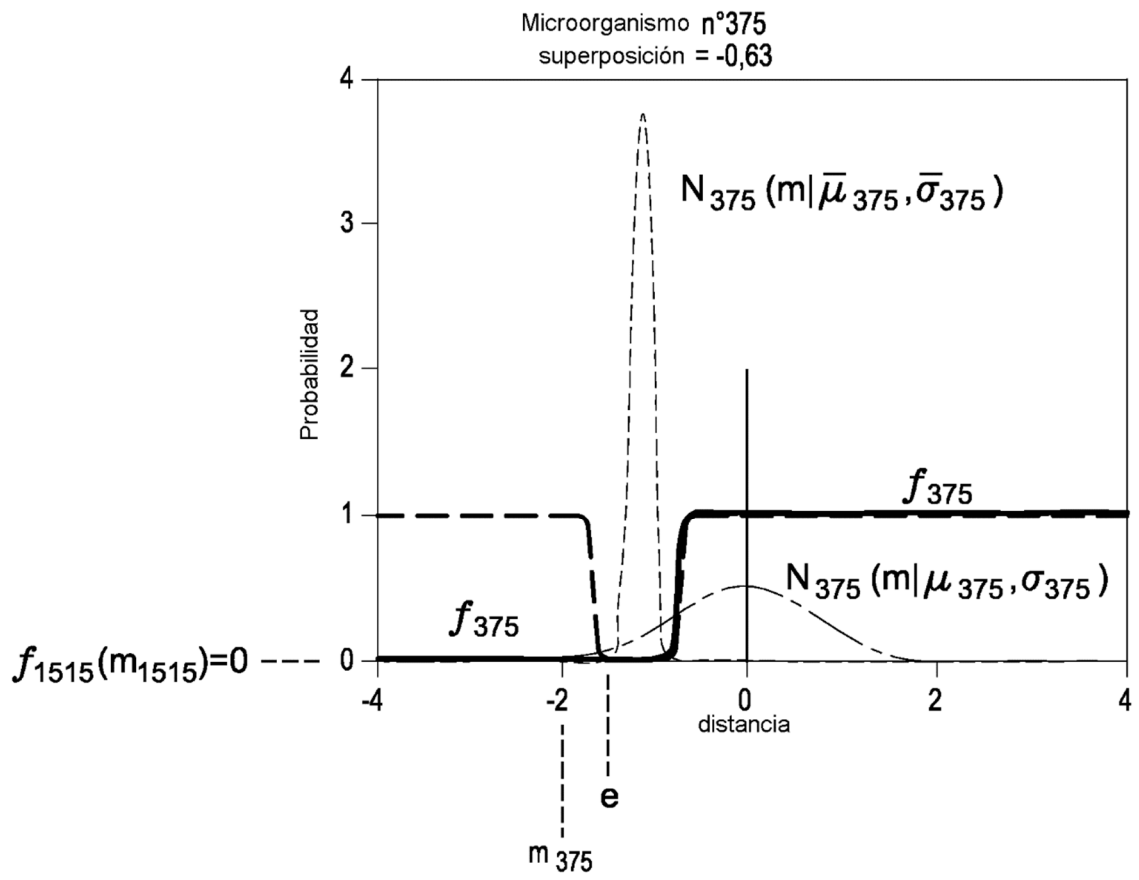
**Fig. 6**



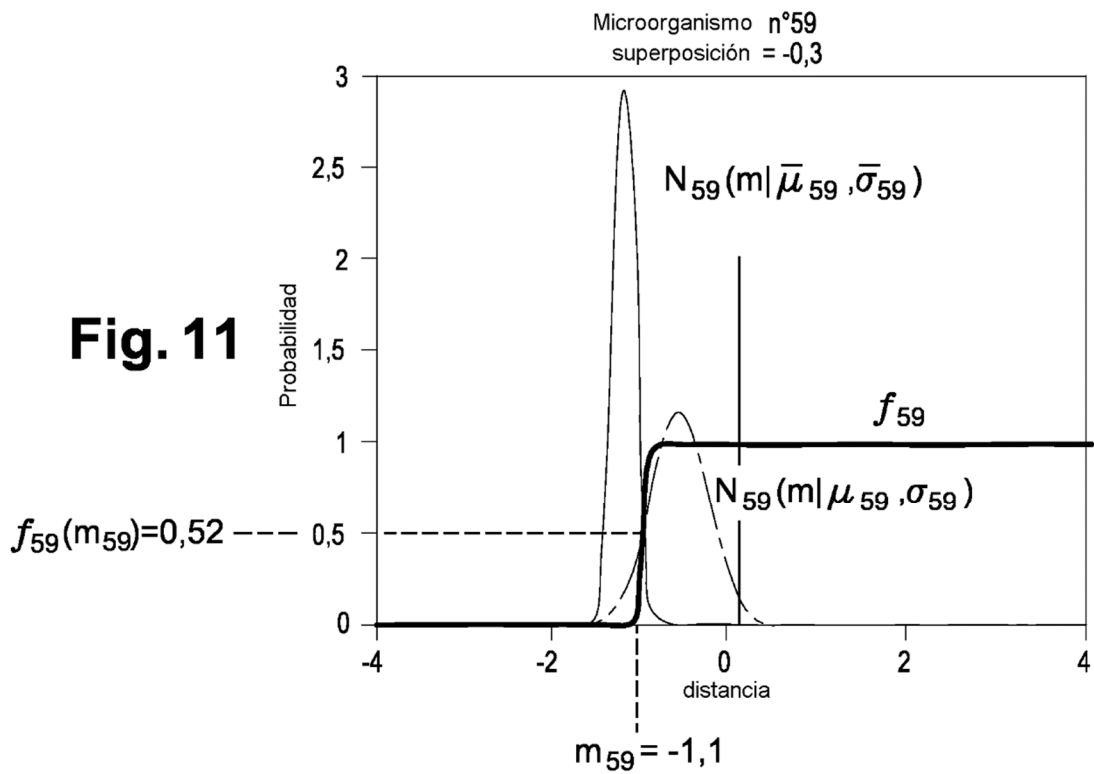
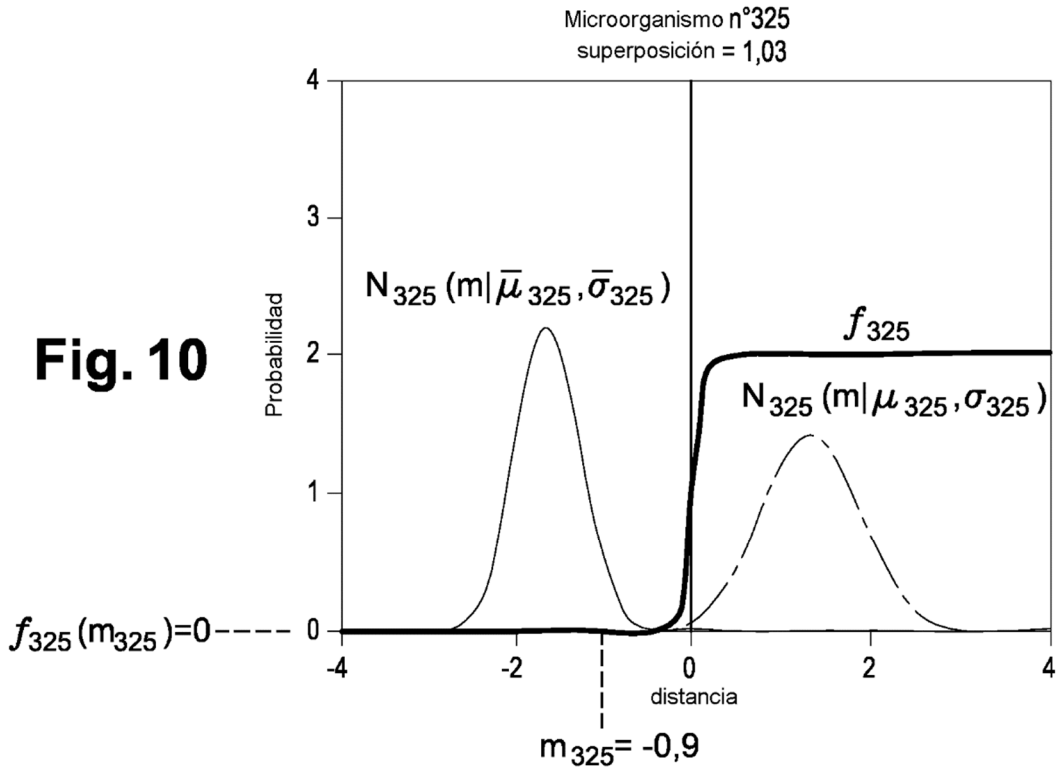
**Fig. 7**

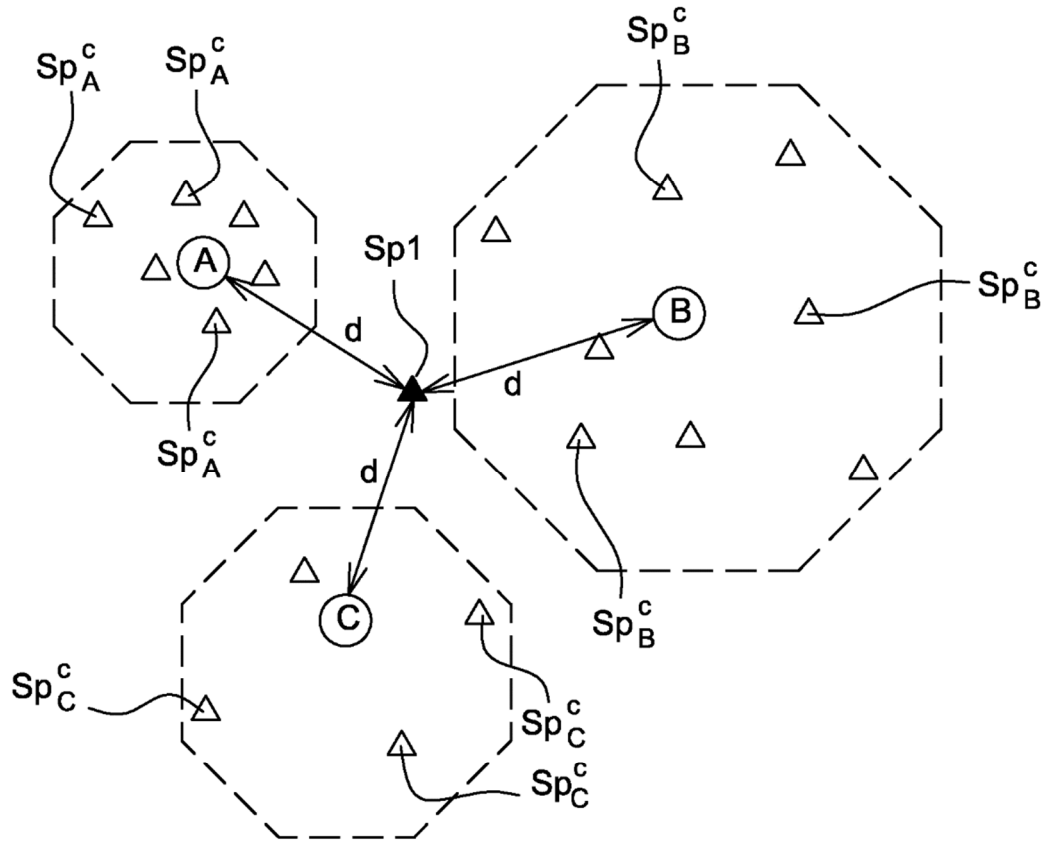


**Fig. 8**



**Fig. 9**





**Fig. 12**