(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 651 163**

(51) Int. CI.:

G06F 19/24 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **28.07.2014 PCT/FR2014/051952**

(87) Fecha y número de publicación internacional: **05.02.2015 WO15015106**

(96) Fecha de presentación y número de la solicitud europea: **28.07.2014 E 14750588 (7)**

(97) Fecha y número de publicación de la concesión europea: **13.09.2017 EP 3028202**

(54) Título: **Procedimiento y dispositivo de análisis de una muestra biológica**

(30) Prioridad:

31.07.2013 FR 1357614

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
24.01.2018

(73) Titular/es:

BIOMÉRIEUX S.A. (100.0%)
69280 Marcy L'Etoile, FR

(72) Inventor/es:

MAHE, PIERRE y
VEYRIERAS, JEAN-BAPTISTE

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Procedimiento y dispositivo de análisis de una muestra biológica

Campo de la invención

La invención se refiere al campo del análisis de muestras biológicas que puede comprender varios microorganismos diferentes, y más particularmente a la detección y a la identificación de mezclas microbianas, a partir de técnicas de medición que producen una señal numérica multidimensional representativa de la muestra biológica objeto del análisis.

Estado de la técnica

Se conoce utilizar la espectrometría o la espectroscopia para identificar unos microorganismos, y más particularmente unas bacterias. Para hacer esto, se prepara una muestra de un microorganismo desconocido a identificar y después se adquiere y se pre-trata un espectro de masa, vibracional o de fluorescencia, de la muestra, en particular para eliminar la línea de base (comúnmente denominada “baseline”) y para eliminar el ruido. El espectro pre-tratado se “compara” entonces con ayuda de una herramienta de clasificación con una base de referencia construida a partir de un conjunto de espectros asociados a taxones de microorganismos identificados, por ejemplo unas especies, mediante un método de referencia.

Más particularmente, la identificación de microorganismos por clasificación consiste clásicamente en una primera etapa de determinación, con la ayuda de un aprendizaje supervisado, de un modelo de clasificación en función de los espectros denominados de “aprendizaje” de microorganismos de los cuales se conoce previamente las especies, definiendo el modelo de clasificación un conjunto de reglas que distinguen estas diferentes especies entre los espectros de aprendizaje, y en una segunda etapa de identificación, o de “predicción”, de un microorganismo particular desconocido. Esta segunda etapa consiste en particular en hacer la adquisición de un espectro del microorganismo a identificar, en pre-tratar el espectro y en aplicar al espectro pre-tratado un modelo de predicción construido a partir del modelo de clasificación a fin de determinar al menos la especie a la que el microorganismo desconocido pertenece.

Típicamente, un aparato de identificación por espectrometría o espectroscopia comprende así un espectrómetro o espectroscopio y una unidad de adquisición y de tratamiento que recibe los espectros medidos, digitalizando estos últimos con el fin de obtener un vector de intensidad numérica multidimensional y utilizando la segunda etapa antes citada en función del vector numérico producido. La primera etapa es, por su parte, realizada por el fabricante del aparato, que determina el modelo de clasificación y el modelo de predicción y lo integra en la máquina antes de su explotación por un cliente.

Hasta ahora, sea cual sea la técnica de medición o el algoritmo de identificación considerado, el análisis de una muestra biológica se limita a muestras que comprenden un solo y único tipo de microorganismo. En efecto, el análisis de muestras biológicas que comprenden una pluralidad de microorganismos diferentes es particularmente difícil y se observa en particular que los algoritmos de predicción que se basan en modelos de clasificación no consiguen detectar que una muestra biológica comprende varios microorganismos, y por lo tanto tampoco identificar los microorganismos contenidos en tal muestra.

Así, previamente a cualquier etapa de identificación por espectrometría o espectroscopia, una muestra a ensayar, de la cual se busca conocer los microorganismos que contiene, se somete en primer lugar a una etapa de tratamiento biológico que tiene como objetivo aislar los diferentes tipos de microorganismos. Se prepara después una muestra biológica objeto de una identificación por espectrometría o espectroscopia a partir de un solo tipo de microorganismo aislado. Por ejemplo, tratándose de la identificación de bacterias, se prepara una solución del producto a ensayar, después la solución obtenida se pone en presencia de uno o varios medios de cultivo, por ejemplo en una o varias cajas de Petri. Después de la incubación, se identifican y se aíslan entonces diferentes colonias bacterianas, pudiendo cada una de ellas ser objeto de una identificación posterior.

Ahora bien, tal preparación de muestra biológica puede llevar mucho tiempo, necesitando en efecto algunos tipos de microorganismos unos tiempos de incubación de varios días. Además, algunos microorganismos necesitan un medio de cultivo muy específico para crecer.

Además del coste que esto genera, existe siempre un riesgo de no hacer crecer todos los diferentes microorganismos comprendidos en el producto a ensayar, y por lo tanto un riesgo de “perder” un microorganismo. Esta etapa preliminar de preparación, hecho obligatorio debido a la incapacidad de los algoritmos de identificación basados en modelos de clasificación para analizar de manera eficaz unas mezclas polimicrobianas, es por lo tanto una fuente de error importante.

Descripción de la invención

El objetivo de la presente invención es proponer un procedimiento de análisis de muestra biológica que permite analizar una muestra biológica, independientemente del hecho de que ésta comprenda uno o varios microorganismos diferentes, en función de una única medición de la muestra, en particular por espectroscopía, espectrometría, o cualquier tipo de medición que produce un vector de intensidad numérica multidimensional.

Para este propósito, la invención tiene por objeto un procedimiento de detección en una muestra biológica de al menos dos microorganismos que pertenecen a dos taxones diferentes entre un conjunto predeterminado y_j } de un número de K taxones de referencia y_j diferentes, estando cada taxón de referencia y_j representado por un vector de intensidad de referencia predeterminada P_j de un espacio R^p, o "prototipo" obtenido sometiendo al menos una muestra biológica de referencia, que comprende un microorganismo que presenta el taxón de referencia, a una técnica de medición que produce una señal numérica multidimensional representativa de la muestra de referencia, y determinando dicho vector de referencia en función de dicha señal numérica multidimensional, en la que p es superior a 1, comprendiendo el procedimiento:

* la adquisición de una señal numérica multidimensional de la muestra biológica mediante la tecnología de medición;

* la determinación de un vector de intensidad x de R^p en función de la señal numérica multidimensional adquirida;

* la construcción de un conjunto $\{\hat{\gamma}_l\}$ de modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, \hat{\gamma}_0)_l$ modelizando el vector de intensidad x según la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} + \hat{\gamma}_0 I_P$$

expresión en la que:

- $\hat{x}_l$ es un vector de R^p que reconstruye el vector de intensidad x por el modelo $\hat{\gamma}_l$;
- $\hat{\gamma}_0$ es un escalar real y I_p es el vector unidad de R^p;
- $\forall j \in [[1,K]]$, $\hat{\gamma}_j$ es el j^enésimo componente de un vector $\hat{\gamma}$ de R_+^K ;
- $\forall j \in [[1,K]]$, $\tilde{P}_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$ y
- $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente predeterminado;

* la selección de un modelo candidato $\hat{\gamma}_{sel}$ entre el conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l$, solución de un problema según la relación:

$$\hat{\gamma}_{sel} = \underset{\hat{\gamma}_l \in \{\hat{\gamma}_l\}}{argmin}\left(C_v(\hat{\gamma}_l) + C_c(\hat{\gamma}_l)\right)$$

expresión en la que:

- $C_v(\hat{\gamma}_l)$ es un criterio que cuantifica un error de reconstrucción entre el vector de intensidad de la muestra biológica x y la reconstrucción $\hat{x}_l$ del vector de intensidad x por un modelo candidato $\hat{\gamma}_l$; y
- $C_c(\hat{\gamma}_l)$ es un criterio que cuantifica la complejidad de un modelo candidato $\hat{\gamma}_l$;

* y la determinación de la presencia en la muestra biológica de al menos dos microorganismos que pertenecen a unos taxones diferentes del conjunto predeterminado $\{\hat{\gamma}_j\}$ de taxones cuando al menos dos componentes $\hat{\gamma}_j$ del

vector $\hat{\gamma}$ del modelo candidato seleccionado $\hat{\gamma}_{sel}$ son superiores a un valor límite predeterminado estrictamente positivo.

Para este propósito, la invención tiene también por objeto un procedimiento de identificación de microorganismos presentes en una muestra biológica entre un conjunto predeterminado $\{\hat{\gamma}_j\}$ de un número de K taxones de referencia diferentes y_j, estando cada taxón de referencia y_j representado por un vector de intensidad de referencia predeterminada P_j de un espacio R^p obtenido sometiendo al menos una muestra biológica de referencia que comprende un microorganismo que presenta el taxón de referencia a una técnica de medición que produce una señal numérica multidimensional representativa de la muestra de referencia y determinando dicho vector de referencia en función de dicha señal numérica multidimensional, en la que p es superior a 1, comprendiendo el procedimiento:

▪ la adquisición de una señal numérica multidimensional de la muestra biológica mediante la tecnología de medición;

▪ la determinación de un vector de intensidad x de R^p en función de la señal numérica multidimensional adquirida;

▪ la construcción de un conjunto $\{\hat{\gamma}_l\}$ de modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, \hat{\gamma}_0)_l$ que modelizan el vector de intensidad x según la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} + \hat{\gamma}_0 I_P$$

expresión en la que:

◦ $\hat{x}_l$ es un vector de R^p que reconstruye el vector de intensidad x por el modelo $\hat{\gamma}_l$;

◦ $\hat{\gamma}_0$ es un escalar real y I_P es el vector unidad de R^p;

◦ $\forall j \in [[1,K]]$, $\hat{\gamma}_j$ es el j^{enésimo} componente de un vector $\hat{\gamma}$ de R_+^K;

◦ $\forall j \in [[1,K]]$, $P_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$; y

◦ $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente predeterminado;

▪ la selección de un modelo candidato $\hat{\gamma}_{sel}$ entre el conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l$, solución de un problema según la relación:

$$\hat{\gamma}_{sel} = \underset{\hat{\gamma}_l \in \{\hat{\gamma}_l\}}{argmin}\left(C_v(\hat{\gamma}_l) + C_c(\hat{\gamma}_l)\right)$$

expresión en la que:

◦ $C_v(\hat{\gamma}_l)$ es un criterio que cuantifica un error de reconstrucción entre el vector de intensidad de la muestra biológica x y la reconstrucción $\hat{x}_l$, del vector de intensidad x por un modelo candidato $\hat{\gamma}_l$; y

◦ $C_c(\hat{\gamma}_l)$ es un criterio que cuantifica la complejidad de un modelo candidato $\hat{\gamma}_l$;

▪ y la determinación de la presencia en la muestra biológica de un microorganismo de taxón y_j del conjunto predeterminado $\{y_j\}$ para cada componente $\hat{\gamma}_j$ del vector $\hat{\gamma}$ del modelo candidato seleccionado $\hat{\gamma}_{sel}$ superior a un valor límite predeterminado estrictamente positivo.

Para este propósito, la invención tiene también por objeto un procedimiento de determinación de la abundancia relativa en una muestra biológica que pertenece a dos taxones diferentes entre un conjunto predeterminado $\{y_j\}$ de un número de K taxones de referencia y_j diferentes, estando cada taxón de referencia y_j representado por un vector de intensidad de referencia predeterminada P_j de un espacio R^p obtenido sometiendo al menos una muestra biológica de referencia, que comprende un microorganismo que presenta el taxón de referencia, a una técnica de medición que produce una señal numérica multidimensional representativa de la muestra de referencia y determinando dicho vector de referencia en función de dicha señal numérica multidimensional, en la que p es superior a 1, comprendiendo el procedimiento:

▪ la adquisición de una señal numérica multidimensional de la muestra biológica mediante la tecnología de medición;

▪ la determinación de un vector de intensidad x de R^p en función de la señal numérica multidimensional adquirida;

▪ la construcción de un conjunto $\{\hat{\gamma}_l\}$ de modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, \hat{\gamma}\,0)_l$ que modelizan el vector de intensidad x según la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} + \hat{\gamma}_0 I_P$$

expresión en la que:

◦ $\hat{x}_l$ es un vector de R^p que reconstruye el vector de intensidad x por el modelo $\hat{\gamma}_l$;

◦ $\hat{\gamma}_0$ es un escalar real y I_p es el vector unidad de R^p;

◦ $\forall j \in [[1,K]]$, $\hat{\gamma}_j$ es el $j^{\text{énésimo}}$ componente de un vector $\hat{\gamma}$ de R_+^K;

◦ $\forall j \in [[1,K]]$, $P_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$; y

◦ $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente predeterminado;

▪ la selección de un modelo candidato $\hat{\gamma}_{sel}$ entre el conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l$, solución de un problema según la relación:

$$\hat{\gamma}_{sel} = \underset{\hat{\gamma}_l \in \{\hat{\gamma}_l\}}{argmin}\left(C_v(\hat{\gamma}_l) + C_c(\hat{\gamma}_l)\right)$$

expresión en la que:

◦ $C_v(\hat{\gamma}_l)$ es un criterio que cuantifica un error de reconstrucción entre el vector de intensidad de la muestra biológica x y la reconstrucción $\hat{x}_l$ del vector de intensidad x por un modelo candidato $\hat{\gamma}_l$; y

◦ $C_c(\hat{\gamma}_l)$ es un criterio que cuantifica la complejidad de un modelo candidato $\hat{\gamma}_l$;

▪ y la determinación de la abundancia relativa en la muestra biológica relativa C_j de un taxón de referencia y_j según la relación:

$$C = J(\hat{\gamma}_{sel})$$

expresión en la que J es una función matricial de $R_+^p \times R_+^K$ en R_+^K y $C = (C_1 \ldots C_j \ldots C_K)^T$ es un vector de R_+^K con $\forall j \in [[1,K]]$, C_j es la abundancia relativa del taxón de referencia y_j.

Por “detección” se entiende aquí la determinación del carácter polimicrobiano de una muestra biológica. La “identificación” de un microorganismo se refiere a la determinación de un dato propio al microorganismo, por ejemplo su especie, su sub-especie, su género, su gram, etc. y de manera más general cualquier dato considerado útil que entra en la construcción de una identidad única del microorganismo.

El término “taxón” designa en particular una noción más amplia que el término “taxón” utilizado para caracterizar la posición de un nudo, de una hoja o de la raíz de una clasificación taxonómica de lo vivo. En los términos de la invención, el término taxón designa cualquier tipo de clasificación de seres vivos que se considere útil. En particular, la invención se aplica a clasificaciones taxonómicas clásicas, clasificaciones basadas en fenotipos clínicos y clasificaciones híbridas basadas en características taxonómicas en el sentido clásico del término y de los fenotipos clínicos.

Por “técnica de medición” se entiende aquí una medición que comprende la producción de una señal compleja que es digitalizada. Entre este tipo de medición, se puede citar por ejemplo la espectrometría de masa, en particular la espectrometría MALDI-TOF y la espectrometría ESI-MS, la espectroscopia vibracional, en particular la espectroscopia RAMAN, la espectroscopia por fluorescencia, en particular la espectroscopia por fluorescencia intrínseca, o la espectroscopia infrarroja. Cada una de estas técnicas produce un espectro que está digitalizado, dando así lugar a una señal numérica multidimensional representativa de la muestra objeto de la medición.

En otras palabras, la invención consiste en producir unos modelos candidatos obtenidos mezclando unos vectores de intensidad representativos cada uno de un taxón previamente identificado con la ayuda de la técnica de medición en cuestión, después en retener el modelo candidato que ofrece el mejor compromiso entre la aproximación del vector de intensidad de la muestra sometida al análisis y la complejidad del modelo candidato. En efecto, se observa que el modelo que estima más fielmente la muestra biológica no es el que permite la reconstrucción más precisa del vector de intensidad, sino el que es al mismo tiempo suficientemente preciso y de complejidad moderada. Los inventores han señalado así que un algoritmo que presenta tal estructura permite al mismo tiempo detectar la presencia de varios microorganismos en una muestra e identificar los microorganismos presentes en la muestra con un porcentaje de éxitos elevado.

Según un modo de realización de la invención, $\forall(i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente de similitud entre los vectores de referencia P_i y P_j de los taxones de referencia y_i y y_j. En particular, los coeficientes de similitud se pueden definir como los productos escalares entre los vectores de intensidad, normalizados o no o, cuando los vectores de referencia enumeran unos picos comprendidos en unos espectros producidos por la técnica de medición, como sus coeficientes de Jaccard. En efecto, se observa que la proximidad biológica entre dos taxones diferentes induce una proximidad entre los dos vectores de referencia de estos taxones. Se puede identificar así un microorganismo de taxón de referencia y_j en una muestra biológica además de, o en lugar de, un microorganismo de taxón de referencia y_i cuyo vector de referencia P_i es muy próximo del vector de referencia P_j del taxón y_j. La creación de vectores de referencia ajustados $P_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$ teniendo en cuenta la proximidad biológica entre unos taxones de referencia minimiza por lo tanto los errores de detección y de identificación.

Según un modo de realización, $\hat{\gamma}_0 = 0$, y la construcción del conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, 0)_l$ comprende la resolución de un conjunto de problemas de optimización para valores de un parámetro λ de R_+, estando cada problema definido según la relación:

$$\hat{\gamma}(\lambda) = \underset{\gamma \in R_+^K}{\operatorname{argmin}} \left(\left\| x - \sum_{j=1}^{K} \gamma_j P_j^{(a)} \right\|^2 + \lambda |\gamma|_1 \right)$$

expresión en la que $|\gamma|_1$ es la norma L1 del vector γ.

En otras palabras, la construcción de los modelos candidatos comprende una penalización de tipo LASSO que hace intervenir un primer término que comprende el error de construcción $x - \sum_{j=1}^{K} \gamma_j P_j^{(a)}$ y un segundo término de ponderación $|\gamma|_1$ en base a la norma L1. Para un término λ nulo, el modelo candidato obtenido es aquel que minimiza el error de reconstrucción bajo una restricción de positividad de los coeficientes del modelo. Como se ha dicho anteriormente, este modelo generalmente no es el que estima mejor la muestra biológica, ya que presenta habitualmente la complejidad más elevada debido a unos componentes $\hat{\gamma}_j$ en mayoría, incluso e totalidad, no nula.

A medida que el parámetro λ aumenta, se observa que los componentes $\hat{\gamma}_j$ se vuelven nulos por un lado, unos

después de otros. Recorriendo los valores de λ, se obtiene por lo tanto un conjunto de modelos candidatos que tienen cada uno una estructura de $\hat{\gamma}$ única. La aplicación de este tipo de algoritmo permite, por lo tanto, realizar una preselección de un número reducido de estructuras de modelos entre las 2^K estructuras de modelo posible. Como, por otro lado, cada problema de optimización es convexo, es posible calcular muy rápidamente los modelos candidatos. Se obtiene así una aceleración sensible del procedimiento según la invención.

En una variante, el escalar γ_0 es no nulo, y el problema de optimización descrito anteriormente se reescribe según la relación:

$$(\hat{\gamma}(\lambda), \hat{\gamma}_0(\lambda)) = \underset{\gamma \in R_+^K, \gamma_0 \in R_+}{\operatorname{argmin}} \left(\left\| x - \left(\sum_{j=1}^{K} \gamma_j P_j^{(a)} + \gamma_0 I_P \right) \right\|^2 + \lambda |\gamma|_1 \right)$$

Se pueden seleccionar así unas estructuras diferentes.

Según otro modo de realización, $\gamma_0 = 0$, y la construcción de $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, 0)_l$ comprende la resolución de un conjunto de problemas de optimización para unos valores de parámetros λ y β de R_+, estando cada problema definido según la relación:

$$\hat{\gamma}(\lambda, \beta) = \underset{\gamma \in R_+^K}{\operatorname{argmin}} \left(\left\| x - \sum_{j=1}^{K} \gamma_j P_j^{(a)} \right\|^2 + \lambda |w_1 \odot \gamma|_1 + \beta |w_2 \odot \gamma|_2 \right)$$

expresión en la que:

- $|\ |_1$ es la norma L1;
- $|\ |_2$ es la norma L2;
- $a \odot b$ es el producto término por término de los vectores a y b; y
- w_1 y w_2 son unos vectores de peso predeterminadas de R_+^K.

En otras palabras, la construcción de los modelos candidatos por el método LASSO se lleva a cabo con la ayuda de un algoritmo de tipo LARS-EN (por «Least Angle Regression - Elastic Net) con penalización de tipo «*elastic net*" (β = 0) combinado con una penalización adaptativa ($w_1 = w_2 = I_K$). El algoritmo LARS-EN es, por ejemplo, aquel de Zou y Hastie que está comprendido en el módulo «*R elasticNet*» disponible en la dirección http://cran.r-project.org/web/packages/elasticnet/. En una variante, sólo se utiliza la penalización de tipo «*elastic net*», es decir que β está colocada nula, o sólo se utiliza la penalización de tipo LASSO adaptativo, es decir que w_1 y w_2 se colocan iguales al vector unidad I_K de R^K. Se pueden obtener unas estructuras diferentes para los modelos candidatos. De manera ventajosa, la versión adaptativa permite incluir información, a priori sobre los taxones que son susceptibles de estar contenidos en la muestra biológica. Por ejemplo, seleccionando un componente de los vectores w_1 w_2 más bajo que los otros componentes, permite hacer más verdadera la presencia del taxón que corresponde a este componente en la muestra biológica.

En una variante, el escalar $\hat{\gamma}_0$ es no nulo, y el problema de optimización descrito anteriormente se reescribe según la relación:

$$(\hat{\gamma}(\lambda, \beta), \hat{\gamma}_0(\lambda, \beta)) = \underset{\gamma \in R_+^K, \gamma_0 \in R_+}{\operatorname{argmin}} \left(\left\| x - \left(\sum_{j=1}^{K} \gamma_j P_j^{(a)} + \gamma_0 I_P \right) \right\|^2 + \lambda |w_1 \odot \gamma|_1 + \beta |w_2 \odot \gamma|_2 \right)$$

Se pueden considerar otros enfoques que permiten una preselección diferente de un algoritmo de tipo LARS-EN, como por ejemplo un algoritmo de tipo «*stepwise*» simple o estructurado, como por ejemplo el descrito en el documento "Structured, sparse regression with application to HIV drug resistance" de Daniel Percival *et al.*, Annals of Applied Statistics, 2011, vol. 5, Nº 2A, 628-644, incluso un enfoque exhaustivo que tiene como objetivo evaluar un número importante de estructuras de modelo candidato entre las 2^K estructuras posibles.

De manera ventajosa, para cada vector $\hat{\gamma}$ solución de un problema de optimización, se calcula un nuevo modelo candidato $\hat{\gamma}_l = \left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right)_l$, y sustituye el modelo $\hat{\gamma}_l = (\hat{\gamma}, 0)_l$ que corresponde al vector $\hat{\gamma}$, correspondiendo los componentes del vector $\hat{\gamma}^{lm}$ del nuevo modelo $\hat{\gamma}_l = \left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right)_l$, a los componentes nulos del vector $\hat{\gamma}$, forzándose a cero, y calculándose el nuevo modelo $\hat{\gamma}_l = \left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right)_l$ resolviendo el problema de optimización según las relaciones:

$$\left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right) = \underset{\substack{\gamma_0^{lm} \in R_+ \\ \gamma^{lm} \in R_+^K}}{\operatorname{argmax}} \left(-\frac{p}{2} ln(2\pi\sigma(x_l)^2) - \frac{1}{2\sigma(x_l)^2} \sum_{b=1}^{p} (x_b - x_{lb})^2 \right)$$

$$\sigma(x_l)^2 = \frac{1}{p} \sum_{b=1}^{p} (x_b - x_{lb})^2$$

$$x_l = \gamma_0^{lm} I_p + \sum_{j:\hat{\gamma}_j \neq 0} \gamma_j^{lm} P_j^{(a)}$$

expresiones en las que:

- x_b es el b^{enésimo} componente del vector de intensidad de la muestra biológica *x*; y
- x_{lb} es el b^{enésimo} componente del vector de reconstrucción $x_l = \gamma_0^{lm} I_p + \sum_{j:\hat{\gamma}_j > 0} \gamma_j^{lm} P_j^{(a)}$.

En otras palabras, los modelos candidatos se recalculan mediante un modelo lineal estándar conservando las estructuras de vectores obtenidas al final de la realización del enfoque LASSO o análogo. Debido a la ponderación del error de reconstrucción por un término en base a la norma L1, los modelos candidatos obtenidos por un enfoque LASSO presentan una verosimilitud reducida, a pesar de que presentan unas estructuras pertinentes. Los modelos candidatos se recalculan ventajosamente conservando las estructuras determinadas por el enfoque LASSO y maximizando su verosimilitud, tal como se define en un modelo lineal estándar. Se refuerza así la calidad del análisis de la muestra biológica, ya que la selección de los modelos candidatos y la estimación de sus efectos se resalizan en dos etapas distintas.

Según un modo de realización, el criterio $C_v(\hat{\gamma}_l)$ que cuantifica el error de reconstrucción es un criterio de verosimilitud. Más particularmente:

$$C_v(\hat{\gamma}_l) = -\frac{p}{2} ln(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{b=1}^{p} (x_b - \hat{x}_{lb})^2$$

expresión en la que:

- $\hat{\sigma}^2 = \frac{1}{p} \sum_{b=1}^{p} (x_b - \hat{x}_{lb})^2$;
- x_b es el b^{enésimo} componente del vector de intensidad de la muestra biológica *x*; y
- $\hat{x}_{lb}$ es el b^{enésimo} componente del vector de reconstrucción $\hat{x}_l$ del modelo candidato $\hat{\gamma}_l$.

Según un modo de realización, el criterio $C_c(\hat{\gamma}_l)$ que cuantifica la complejidad del modelo $\hat{\gamma}_l$ cuantifica dicha complejidad en términos de número de componentes $\hat{\gamma}_j$ del vector $\hat{\gamma}$ estrictamente positivas. Más particularmente:

$$\text{Si } \hat{\gamma}_0 = 0 \quad \text{entonces} \quad C_c(\hat{\gamma}_l) = \left(1 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) \ln p$$

$$\text{Si } \hat{\gamma}_0 \neq 0 \quad \text{entonces} \quad C_c(\hat{\gamma}_l) = \left(2 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) \ln p$$

expresiones en las que la función 1(.) es igual a 1 si su argumento es verdadero y cero sino lo es.

En una variante, cuando el escalar $\hat{\gamma}_0$ está colocado igual a cero durante el cálculo de los modelos candidatos, y por lo tanto el escalar γ_0^{lm}, el criterio $C_c(\hat{x}_l)$ se reescribe según la relación:

$$C_c(\hat{\gamma}_l) = \left(1 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) \ln p$$

Así, según un modo de realización preferido, el modelo candidato $\hat{\gamma}_{sel}$ seleccionado es el que minimiza la función según la relación:

$$-\frac{p}{2} ln(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{b=1}^{p} (x_b - \hat{x}_{lb})^2 + \left(2 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) ln\, p$$

o la función según la relación:

$$-\frac{p}{2} ln(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{b=1}^{p} (x_b - \hat{x}_{lb})^2 + \left(1 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) ln\, p$$

En otras palabras, el modelo candidato seleccionado es el que minimiza un criterio “BIC” (acrónimo de *“Bayesian Information Criterion*”) que propone una selección eficaz del modelo. Se podrá, por ejemplo, hacer referencia al documento “El criterio BIC: fondements théoriques y interprétation”, de Emilie Labarbier y Tristant Mary-Huard, INRIA, Informe de investigación n°5315, septiembre de 2004, para una descripción más detallada de este criterio.

No obstante, son posibles otros criterios de selección, como por ejemplo un criterio “AIC” (acrónimo de *“Akaike Information Criterion*”), “MLD” (acrónimo de *“Minimum Rescription Length*”), “Cp de Mallows”, o de manera general cualquier criterio que combina un criterio de verosimilitud o de error de reconstrucción con un criterio de complejidad.

Según un modo de realización, los taxones pertenecen a un mismo nivel taxonómico, en particular el nivel especie, género o sub-especie. En una variante, los taxones pertenecen a al menos dos niveles taxonómicos diferentes, en particular unas especies, unos géneros, y/o unas sub-especies. En particular, si un grado de similitud entre un conjunto de taxones definidos dentro de un primer nivel taxonómico es superior a un límite predeterminado, entonces para la constitución del conjunto predeterminado $\{y_j\}$ de los taxones de referencia, dichos taxones se agrupan y sustituyen por un taxón de referencia definido a un segundo nivel taxonómico, superior al primer nivel taxonómico.

En otras palabras, el procedimiento según la invención es libre de seleccionar unos niveles de descripción diferentes de los microorganismos. Por ejemplo, es posible combinar unas especies con unos géneros sin que esto plantee problema en particular. Gracias a la invención, es por lo tanto posible seleccionar unos taxones de referencia que se

distinguen suficientemente los unos de los otros frente a vectores de referencia, y así minimizar los errores de detección y de identificación. Por ejemplo, cuando los espectros de especies dentro de un mismo género particular presentan unos grados de similitud muy importantes, arriesgando por lo tanto poner el algoritmo de detección o de identificación en dificultad, es posible seleccionar mejor el género que las especies, y al mismo tiempo preferir el nivel especie para los otros microorganismos.

Según un modo de realización, unos taxones pertenecen a un primer nivel taxonómico, y un nuevo modelo del vector *x* se calcula estimando la contribución de dichos taxones a un segundo nivel taxonómico, superior al primer nivel taxonómico, sumando los componentes del vector $\hat{\gamma}$ relacionados con dicho nivel superior. En particular, el modelo del vector *x* se calcula para el nivel taxonómico superior si un grado de similitud dentro del primer nivel es superior a un límite predeterminado.

En otras palabras, gracias a la invención, es posible identificar el nivel taxonómico superior de un microorganismo gracias a los resultados obtenidos por el algoritmo a nivel taxonómico inferior. Esto permite, por ejemplo, conservar un nivel taxonómico idéntico para todos los microorganismos, incluso cuando los microorganismos presenten una similitud muy elevada a dicho nivel, y compensar los errores de detección y de identificación resultantes calculando un modelo candidato para el nivel taxonómico superior. Este enfoque es también aplicable a unos taxones de referencia considerados a niveles diferentes, pudiendo calcularse unos niveles superiores según la demanda, en particular cuando el modelo candidato finalmente seleccionado comprende unos taxones juzgados muy similares.

Según un modo de realización, la técnica de medición produce un espectro y los vectores de intensidad de referencia P_j son unas listas de picos comprendidos en los espectros de los taxones de referencia y_j. En particular, la técnica de medición comprende una espectrometría de masa.

Según un modo de realización:

$$C_j = \frac{\hat{\gamma}_{j,sel}}{\sum_{i=1}^{K} \hat{\gamma}_{i,sel}}$$

expresión en la que $\forall j \in [[1,K]]$, $\hat{\gamma}_{j,sel}$ es el j^{enésimo} componente del vector $\hat{\gamma}$ del modelo seleccionado $\hat{\gamma}_{sel}$.

La invención tiene también por objeto un dispositivo de análisis de una muestra biológica que comprende:

- un espectrómetro o un espectroscopio apto para producir unos espectros de la muestra biológica;
- una unidad de cálculo apta para realizar un procedimiento del tipo antes citado.

Breve descripción de las figuras

La invención se entenderá mejor a partir de la lectura de la descripción siguiente, dada únicamente a título de ejemplo, y realizada en relación con los dibujos anexos, en los que:

- la figura 1 es un organigrama que ilustra un procedimiento según la invención;
- las figuras 2A y 2B son, respectivamente, una matriz de similitud entre varias especies de bacterias utilizadas para ensayar el procedimiento según la invención y unos vectores de picos de mezclas de dichas bacterias; y
- las figuras 3A y 3B son, respectivamente, unos resultados de la detección y de la identificación según la invención, respectivamente realizadas a nivel especie y a nivel género sobre las mezclas.

Descripción detallada de la invención

Se describirá ahora, en relación con el organigrama de la figura 1 un modo de realización de la invención aplicado a la espectrometría de masa MALDI-TOF (acrónimo de *"Matrix-assisted lasser desorption/ionization time of flight"*) y para un único nivel taxonómico, a saber el nivel especie. La espectrometría de masa MALDI-TOF es bien conocida y no se describirá por lo tanto más en detalle a continuación. Se podrá, por ejemplo, hacer referencia al documento de Jackson O. Lay, "Maldi-tof spectrometry of bacteria", Mass Spectrometry Reviews, 2001, 20, 172-194.

El procedimiento empieza por una etapa 10 de construcción de un conjunto $\{P_j\} = \{P_1\ P_2\ ...\ P_K\}$ de *K* vectores de intensidad de referencia P_j, cada uno asociado a una especie de referencia y_j de microorganismo previamente identificado, y se continúa por una etapa 12 de análisis de una muestra biológica de la cual se busca saber si comprende una o varias especies de referencia diferentes, y/o de la cual se busca identificar la o las especies de

referencia que es susceptible contener, y/o de la cual se busca cuantificar la abundancia de los microorganismos presentes en la muestra.

Un ejemplo de realización de la etapa 10 se describe ahora para una especie de referencia y_j. La etapa 10 comprende la adquisición, en 14**,** de al menos un espectro de masa numérico de la especie y_j en un intervalo de Thomson $[m_{min};m_{max}]$ predeterminado por una espectrometría MALDI-TOF. Por ejemplo, se utilizan varias cepas de un microorganismo que pertenecen a la especie y_j y se adquiere un espectro para cada una de las cepas. Los espectros numéricos adquiridos para la especie y_j se pretratan después, ventajosamente después de una transformación logarítmica, a fin en particular de quitar el ruido de estos y quitar su línea de base, de manera en sí misma conocida.

Se realiza entonces en 16 una identificación de los picos presentes en los espectros adquiridos**,** por ejemplo mediante un algoritmo de detección de picos basado en la detección de valores máximos locales. Se genera así una lista de los picos para cada espectro adquirido, que comprende la localización y la intensidad de los picos del espectro.

Se continúa el procedimiento, en 18**,** mediante una etapa de cuantificación, o "*binning*". Para ello, el intervalo de Thompsons $[m_{min}; m_{max}]$ se subdivide en p intervalos, o "*bins",* de anchuras predeterminadas, por ejemplo idénticas. Cada lista de picos se reduce reteniendo sólo un pico por intervalo, por ejemplo el pico que presenta la mayor intensidad. Se reduce así cada lista a un vector de R^p que tiene como componente la intensidad de los picos retenidos para los intervalos de cuantificación, significando el valor nulo para un componente que no se ha detectado, ni por lo tanto conservado, ningún pico en el intervalo correspondiente. Se produce después un vector numérico multidimensional $P_j \in R^p$*,* también denominado "prototipo", para la especie y_j en función de las listas de picos reducidas. Cada componente del vector P_j se establece particularmente nulo si la frecuencia de los componentes correspondientes de las listas reducidas que son estrictamente positivas es inferior a un límite, por ejemplo un 30%, y si no se selecciona igual al valor medio de los componentes correspondientes de las listas reducidas, que son estrictamente positivas o iguales a la media de los componentes correspondientes de las listas reducidas.

En particular, para la espectrometría MALDI-TOF, $[m_{min}; m_{max}]$ [3000;17000]. En efecto, se ha observado que las informaciones suficientes para la identificación de los microorganismos están agrupadas en este intervalo de relación masa sobre carga, y que por lo tanto no es necesario tener en cuenta un intervalo más ancho. El intervalo $[m_{min}; m_{max}]$ se subdivide en p = 1300 intervalos constantes. El vector P_j es por lo tanto un vector de R^{1300}*.* En una variante, la anchura de los intervalos es creciente de manera logarítmica, como se describe en la solicitud EP 2 600 385.

En una variante, el vector P_j se "binariza" estableciendo el valor de un componente del vector P_j a "1" cuando un pico está presente en el intervalo correspondiente, y a "0" cuando ningún pico está presente en este intervalo. Esto tiene por efecto hacer más firme el análisis de muestra biológica de la etapa 12. Los inventores han notado, en efecto, que la información pertinente, en particular para la identificación de una bacteria, está contenida esencialmente en la ausencia y/o en 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 variabilidad, es difícil tener en cuenta los valores brutos de intensidad en las herramientas de clasificación.

Por supuesto, el vector P_j de la especie y_j se puede obtener mediante cualquier método juzgado útil a fin de producir un vector representativo de la especie y_j. Por ejemplo, los espectros de las cepas de la especie y_j son objeto de un tratamiento estadístico a fin de producir un único espectro. El espectro único es después objeto de una detección de picos y la lista de picos se cuantifica entonces conservando en cada intervalo de la cuantificación sólo el pico de mayor intensidad. El tratamiento estadístico puede, por ejemplo, ser el cálculo de la media de los espectros, el cálculo de un espectro medio, o la selección del espectro que presenta la distancia media a todos los demás espectros de la especie más débil. Asimismo, la etapa de cuantificación 18**,** que permite reducir de manera importante el número de datos a tratar, y garantizar al mismo tiempo una fuerza algorítmica, es opcional. El vector P_j puede, por ejemplo, estar constituido del espectro numérico directamente obtenido después de la etapa 14 de adquisición y de pretratamiento. De manera general, puede ser conveniente cualquier método que permite producir para la especie y_j un vector numérico que comprende una firma única de esta especie.

Los vectores $\{P_j\}$ obtenidos mediante la etapa 10 de construcción se memorizan entonces en una base de datos. La base de datos se incorpora después en un sistema de análisis de muestras biológicas por espectrometría de masa que comprende un espectrómetro de masa, de tipo MALDI-TOF, así como una unidad de tratamiento de informaciones, conectada al espectrómetro y apta para recibir, digitalizar y tratar los espectros de masa adquiridos realizando la etapa de análisis 12. El sistema de análisis puede también comprender una unidad de tratamiento de informaciones distante del espectrómetro de masa. Por ejemplo, el análisis numérico se realiza sobre un servidor distante accesible por un usuario mediante un ordenador conectado a la red internet a la que está también conectado el servidor. El usuario carga unos espectros de masa numéricos no tratados obtenidos por un espectrómetro de masa de tipo MALDI-TOF en el servidor, y este último realiza entonces el algoritmo de análisis y reenvía los resultados del algoritmo al ordenador del usuario. Se señala que la base de datos, en particular la

integrada en el sistema de análisis, se puede actualizar en cualquier momento, en particular para añadir, sustituir y/o retirar un vector de intensidad de referencia.

Se describirá ahora un ejemplo de realización de la etapa de análisis 12 de una muestra biológica de la cual se busca saber si comprende uno o varios tipos de microorganismo y/o de la cual se busca identificar el o los microorganismos que contiene y/o de la cual se busca cuantificar la abundancia relativa de varios microorganismos presentes en la muestra.

La etapa de análisis 12 comprende una primera etapa 20 de preparación de la muestra biológica con la espectrometría MALDI-TOF, en particular la incorporación de la muestra en una matriz, como se conoce en sí misma. Más particularmente, la muestra no soporta ninguna etapa preliminar que tiene como objetivo aislar los diferentes tipos de microorganismos que contiene.

El análisis 12 se prosigue por una etapa 22 de adquisición de un espectro de masa numérico de la muestra biológica por un espectrómetro MALDI-TOF y se le quita el ruido al espectro adquirido y se retira su línea de base.

En una etapa 24 siguiente, se realiza una detección de los picos del espectro numérico y de determinación de un vector de intensidad x de R^p a partir de los picos detectados. Por ejemplo, una cuantificación del espacio de Thomson tal como se ha descrito anteriormente se lleva a cabo conservando sólo el pico de mayor intensidad en un intervalo de cuantificación. De manera general, el vector de intensidad x se puede generar mediante cualquier método apropiado.

Una vez obtenido el vector de intensidad x de R^p en función del espectro de masa de la muestra biológica, se continúa el análisis, en 26, mediante la construcción de un conjunto $\{\hat{\gamma}_l\}$ de modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}_j, \hat{\gamma}_0)_l$ que modelizan el vector de intensidad x según la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} + \hat{\gamma}_0 I_P \qquad (1)$$

expresión en la que:

- $\hat{x}_l$ es un vector de R^p que reconstruye el vector de intensidad x por el modelo $\hat{\gamma}_l$;
- K es el número de vectores de intensidad de referencia P_j memorizados en la base de datos;
- $\hat{\gamma}_0$ es un escalar real y $I_P = (11 \ldots 1)^T$ es el vector unidad de R^p;
- $\forall j \in [[1,K]]$, $\hat{\gamma}_j$ es el j$^{\text{enésimo}}$ componente de un vector $\hat{\gamma}$ de R_+^K;
- $\forall j \in [[1,K]]$, $P_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$; y
- $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente predeterminado;

Más particularmente, los coeficientes a_{ij} son unos coeficientes que cuantifican la similitud, o la "proximidad" entre los vectores de intensidad de referencia P_j, en particular unos coeficientes de Jaccard según la relación:

$$a_{ij} = \frac{N_{ij}^C}{(N_i + N_j) - N_{ij}^C} \qquad (2)$$

en la que N_i es el número de componentes no nulos del vector P_i, N_j es el número de componentes no nulos del vector P_j, y N_{ij}^C es el número de componentes no nulos que comparten los vectores P_i y P_j.

Más particularmente, la construcción 26 del conjunto $\{\hat{\gamma}_I\}$ comprende una primera etapa 28 de selección de un conjunto $\{\tilde{\gamma}\}$ de estructuras $\tilde{\gamma}$ de complejidad crecientes para los vectores $\hat{\gamma}$ de los modelos candidatos $\hat{\gamma}_I$, seguida de una etapa 30 de cálculo de modelos candidatos que tienen las estructuras de $\tilde{\gamma}$ seleccionadas.

En particular, la etapa 28 consiste en seleccionar un conjunto $\{\tilde{\gamma}\}$ de vectores binarios $\tilde{\gamma}$ de R^K que comprende un número creciente de componentes nulos, indicando cada vector $\tilde{\gamma}$ cuales componentes del vector $\hat{\gamma}$ de un modelo candidato $\hat{x}_I$ están libres o forzados a 0. En particular, un componente de valor 0 del vector $\tilde{\gamma}$ indica que el componente correspondiente del vector $\hat{\gamma}$ está forzado a 0, y un componente de valor 1 del vector $\tilde{\gamma}$ indica que el componente correspondiente del vector $\hat{\gamma}$ está libre de tomar un valor positivo no nulo. Por ejemplo, poniendo $p = 3$, y seleccionado un vector $\tilde{\gamma} = (1\ 0\ 1)^T$, se determina que un modelo candidato $\hat{x}_I$ estará calculado teniendo el segundo componente del vector $\hat{\gamma}$ forzado a cero y los primero y tercer componentes del vector $\hat{\gamma}$ libres de tomar unos valores positivos no nulos.

De manera ventajosa, las estructuras $\tilde{\gamma}$ de los vectores $\hat{\gamma}$ se seleccionan realizando un enfoque, o "penalización", LASSO, es decir resolviendo un conjunto de problemas de optimización para unos valores de un parámetro λ de R_+, estando cada problema definido según la relación:

$$(\hat{\gamma}(\lambda), \hat{\gamma}_0(\lambda)) = \underset{\gamma \in R_+^K, \gamma_0 \in R_+}{\operatorname{argmin}} \left(\left\| x - \left(\sum_{j=1}^{K} \gamma_j P_j^{(a)} + \gamma_0 I_P \right) \right\|^2 + \lambda |\gamma|_1 \right) \qquad (3)$$

expresión en la que $|\gamma|_1$ es la norma L1 del vector γ.

En particular partiendo de $\lambda = 0$, que corresponde a un vector $\hat{\gamma}$ (0) del cual cada componente está libre de tomar cualquier valor positivo o nulo, a medida que el parámetro λ crece, la penalización LASSO anula una a una uno de los componentes del vector $\hat{\gamma}$ (λ) hasta alcanzar un vector $\hat{\gamma}$ (λ) nulo. Se obtiene así un número reducido, es decir muy inferior a 2^K, los más frecuentemente próximo o igual a K, vectores $\hat{\gamma}$ (λ) de estructuras diferentes $\tilde{\gamma}$ (λ). Por otro lado, el enfoque LASSO tiene como objetivo minimizar el error de reconstrucción $\left\| x - \left(\sum_{j=1}^{K} \gamma_j P_j^{(a)} + \gamma_0 I_P \right) \right\|$ bajo restricción, cada una de las estructuras seleccionadas representa una estructura pertinente, incluso la mejor estructura, para la complejidad que presenta, es decir el número de sus componentes nulos.

El enfoque LASSO, y sus variantes como la penalización "elastic net", se realiza por ejemplo con la ayuda del algoritmo LARS-EN de Zou y Hastie que está comprendido en el módulo *"R elasticNet"* disponible en la dirección http://cran.r-project.org/web/packages/elasticnet/.

Para cada estructura $\tilde{\gamma}$ seleccionada, la etapa 30 de cálculo del modelo candidato $\hat{\gamma}_I$ que tiene un vector $\hat{\gamma}$ según la estructura $\tilde{\gamma}$ consiste de manera preferida en maximizar un criterio de verosimilitud entre el vector de reconstrucción $\hat{x}_I$ del modelo $\hat{\gamma}_I$ y el vector de intensidad x de la muestra biológica. En particular, el modelo candidato $\hat{\gamma}_I = (\hat{\gamma}, \hat{\gamma}_0)$ se calcula resolviendo el problema de optimización según las relaciones:

$$(\hat{\gamma} \quad , \hat{\gamma}_0 \quad) = \underset{\substack{\gamma_0 \in R_+ \\ \gamma \in R_+^K}}{argmax} \left(-\frac{p}{2} ln(2\pi\sigma(x_l)^2) - \frac{1}{2\sigma(x_l)^2} \sum_{b=1}^{p} (x_b - x_{lb})^2 \right) \qquad (4)$$

$$\sigma(x_l)^2 = \frac{1}{p} \sum_{b=1}^{p} (x_b - x_{lb})^2 \qquad (5)$$

$$x_l = \gamma_0 I_p + \sum_{j=1}^{K} (\tilde{\gamma} \odot \gamma)_j \, P_j^{(a)} \qquad (6)$$

expresiones en las que:

- $a \odot b$ es el producto término a término de los vectores a y b;
- x_b es el b^{enésimo} componente del vector de intensidad de la muestra biológica x; y
- x_{lb} es el b^{enésimo} componente del vector de reconstrucción x_l.

De manera equivalente, cuando las estructuras $\tilde{\gamma}$ se determinan por el enfoque LASSO de la relación (3), el modelo candidato $\hat{\gamma}_l = (\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm})_l$ tiene los componentes de su vector $\hat{\gamma} = \hat{\gamma}^{lm}$ forzados a 0 cuando los componentes correspondientes del vector $\tilde{\gamma}$ son iguales a 0, lo que corresponde a un vector de reconstrucción $\hat{x}_l$ que se reescribe $\hat{x}_l = \hat{\gamma}_0^{lm} I_p + \sum_{j:\hat{\gamma}_j>0} \hat{\gamma}_j^{lm} P_j^{(a)}$.

El problema de optimización de las relaciones (4), (5), (6) se reescribe entonces según las relaciones:

$$(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}) = \underset{\substack{\gamma_0^{lm} \in R_+ \\ \gamma^{lm} \in R_+^K}}{argmax} \left(-\frac{p}{2} ln(2\pi\sigma(x_l)^2) - \frac{1}{2\sigma(x_l)^2} \sum_{b=1}^{p} (x_b - x_{lb})^2 \right) \qquad (4bis)$$

$$\sigma(x_l)^2 = \frac{1}{p} \sum_{b=1}^{p} (x_b - x_{lb})^2 \qquad (5)$$

$$x_l = \gamma_0^{lm} I_p + \sum_{j:\hat{\gamma}_j \neq 0} \gamma_j^{lm} P_j^{(a)} \qquad (6bis)$$

expresiones en las que j: $\hat{\gamma}_j \neq 0$ significa los componentes j del vector $\hat{\gamma}$ calculado por el enfoque LASSO que son no nulos, por lo tanto estrictamente positivos.

Una vez el conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l$ calculado, la etapa 12 de análisis de la muestra biológica se prosigue por una etapa 32 de selección de un modelo candidato $\hat{\gamma}_{sel} = (\hat{\gamma}_{sel}, \hat{\gamma}_{0sel})$ entre el conjunto $\{\hat{\gamma}_l\}$, siendo el modelo candidato seleccionado $\hat{\gamma}_{sel}$ el considerado como estimando de manera más pertinente el vector de intensidad x de la muestra biológica analizada.

Más particularmente, la selección del modelo candidato $\hat{\gamma}_{sel}$ consiste en seleccionar el modelo que ofrece el mejor compromiso entre la aproximación del vector x y la complejidad de la estructura del modelo. Para hacer esto, el

modelo $\hat{\gamma}_{sel}$ es el que minimiza un criterio mezclando un criterio $C_v(\hat{\gamma}_l)$ que cuantifica el error de reconstrucción de la estimación, o la reconstrucción del vector *x* y un criterio $C_c(\hat{\gamma}_l)$ que cuantifica la complejidad de la estimación, y en particular el número de componentes del vector $\hat{\gamma}$ no nulos. Ventajosamente, el modelo $\hat{\gamma}_{sel}$ se selecciona minimizando un criterio "BIC", siendo el modelo $\hat{\gamma}_{sel}$ la solución del problema de optimización según la relación:

$$\hat{\gamma}_{sel} = \underset{\hat{\gamma}_l \in \{\hat{\gamma}_l\}}{argmin}\left(C_v(\hat{\gamma}_l) + C_c(\hat{\gamma}_l)\right) \quad (7)$$

$$C_v(\hat{\gamma}_l) = -\frac{p}{2} ln(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{b=1}^{p}(x_b - \hat{x}_{lb})^2 \quad (8)$$

$$C_c(\hat{\gamma}_l) = \left(2 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) ln\, p \quad (9)$$

expresiones en las que la función 1(.) es igual a 1 si su argumento es verdadero y cero si no lo es y $\sigma^2 = \sigma(\hat{x}_l)^2$.

Así, los modelos candidatos $\hat{\gamma}_l$ calculados maximizando el criterio de verosimilitud según la relación (5), maximizan también el criterio de verosimilitud de la relación (8). Por otro lado, la complejidad de los modelos candidatos $\hat{\gamma}_l$ hace intervenir el número de los componentes de sus vectores $\hat{\gamma}$ no nulos. Se observa que la selección con la ayuda de este tipo de criterio es firme y pertinente. En particular, el modelo finalmente seleccionado $\hat{\gamma}_{sel}$ es aquel que enumera de manera pertinente las especies y_j, representadas respectivamente por los componentes $\hat{\gamma}_j$ del vector $\hat{\gamma}$.

La etapa de análisis 12 se prosigue entonces por la explotación, en 34, del modelo seleccionado $\hat{\gamma}_{sel} = (\hat{\gamma}_{sel}, \hat{\gamma}_{0sel})$ para deducir unas informaciones sobre la muestra biológica analizada.

Más particularmente, se realiza al menos una de las explotaciones siguiente:

a. la muestra biológica se determina como comprendiendo varias especies de referencia diferentes de microorganismos si el número de componentes del vector $\hat{\gamma}_{sel}$, que son superiores a un valor límite predeterminado positivo o nulo, es superior o igual a 2;

b. el número de especies de referencia diferentes en la muestra biológica se calcula como igual al número de componentes del vector $\hat{\gamma}_{sel}$ superiores a un valor límite predeterminado positivo o nulo;

c. una especie de referencia y_j se identifica en la muestra biológica cuando el $j^{enésimo}$ componente del vector $\hat{\gamma}_{sel}$ es superior a un valor límite positivo o nulo predeterminado, ajustando el valor límite la sensibilidad de la identificación;

d. la abundancia relativa C_j de una especie y_j en la muestra biológica se calcula según la relación:

$$C_j = \frac{\hat{\gamma}_{selj}}{\sum_{i=1}^{K}\hat{\gamma}_{seli}} \quad (10)$$

en la que $\hat{\gamma}_{selj}$ es el $j^{enésimo}$ componente del vector $\hat{\gamma}_{sel}$;

e. los resultados que se refieren a unas especies de referencia que pertenecen a un mismo nivel taxonómico superior, en particular el género de éstas, se agrupan calculando un escalar $\hat{\gamma}_s^{sup}$ igual a la suma de los componentes del vector $\hat{\gamma}_{sel}$ que corresponde a dichas especies. El nivel taxonómico superior se identifica después si el escalar $\hat{\gamma}_s^{sup}$ es superior al valor límite predeterminado positivo o nulo. En particular, el j[enésimo] componente del vector $\hat{\gamma}_{sel}$ que corresponde a cada una de las especies que pertenecen al nivel superior puede ser inferior al valor límite, caso en el que las especies no están identificadas en la mezcla, mientras existen en realidad en la muestra biológica al menos una especie que pertenece al nivel superior. Por ejemplo, cuando en un mismo género que reagrupa unas especies de referencias y_1, y_2, y_3 difícilmente diferenciables por la espectrometría de masa, la presencia únicamente de la especie de referencia particular y_1 en la muestra biológica puede llevar, no a un vector $\hat{\gamma}_{sel}$ del cual el componente que corresponde a la especie y_1 es no nulo y cuyos componentes que corresponden a las especies y_2 y y_3 son nulos, sino a un vector $\hat{\gamma}_{sel}$ cuyos componentes que corresponden a las especies y_1, y_2, y_3 son todos no nulos, inferiores al valor límite. Sumando los componentes de estas especies, se obtiene un valor que supera el límite y por lo tanto una detección del género en la mezcla biológica. En una variante, o de manera suplementaria, calculando el escalar $\hat{\gamma}_s^{sup}$ para el nivel superior, la abundancia relativa del nivel superior se calcula según la relación $C_{sup} = \frac{\hat{\gamma}_s^{sup}}{\sum_{i=1}^{K} \hat{\gamma}_{seli}}$.

El reagrupamiento se realiza de manera ventajosa automáticamente, en particular cuando se observa que las especies de un mismo nivel taxonómico, por ejemplo las especies que pertenecen a un mismo género, son muy similares en términos de espectro de masa. En particular, la similitud de las especies se calcula, por ejemplo, mediante coeficientes de Jaccard de su vector de intensidad de referencia, y si la similitud calculada es superior a un valor límite, entonces el reagrupamiento de los resultados de las especies se realiza automáticamente.

Los resultados de la explotación se memorizan entonces en una memoria informática, por ejemplo la del dispositivo de análisis y/o mostrados en una pantalla destinada al usuario.

Se ha descrito un modo de realización particular de la invención. No obstante, son posibles numerosas variantes, en particular las variantes siguientes consideradas solas o en combinación.

Según una variante, los modelos candidatos $\hat{\gamma}_I$ no comprenden término $\hat{\gamma}_0 I_P$ durante de la selección por el enfoque LASSO. La relación (1) se reescribe entonces:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} \qquad (1\text{bis})$$

Según una variante, los modelos candidatos recalculados $\hat{\gamma}_I$ durante la etapa 30, es decir los utilizados para la selección del modelo final $\hat{\gamma}_{sel}$, no comprenden términos $\hat{\gamma}_0 I_P$. Las relaciones (4) a (11) se deducen fácilmente de esta simplificación. En particular, conviene señalar que la relación (10) se reescribe según la relación:

$$C_c(\hat{\gamma}_l) = \left(1 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) \ln p \qquad (9\text{bis})$$

Según una variante, los coeficientes a_{ij} valen 1 cuando $i = j$ y valen 0 cuando $i \neq j$, en tal caso la relación (1) se reduce a la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j \quad + \hat{\gamma}_0 I_P \qquad (1ter)$$

Según una variante, el coeficiente de similitud a_{ij} entre dos vectores de intensidad de referencia P_i y P_j es el producto escalar de estos.

Según una variante, la selección de las estructuras $\tilde{\gamma}$ de los vectores $\hat{\gamma}$ de los modelos candidatos $\hat{\gamma}_l$ se realiza realizando unos algoritmos derivados del enfoque LASSO de la relación (3), en particular un problema de optimización según una de las relaciones siguientes:

$$(\hat{\gamma}(\lambda), \hat{\gamma}_0(\lambda)) = \underset{\gamma \in R_+^K, \gamma_0 \in R_+}{argmin} \left(\left\| x - \left(\sum_{j=1}^{K} \gamma_j P_j^{(a)} + \gamma_0 I_P \right) \right\|^2 + \lambda \left| w_1 \odot \gamma \right|_1 \right) \qquad (3bis)$$

$$(\hat{\gamma}(\lambda, \beta), \hat{\gamma}_0(\lambda, \beta)) = \underset{\gamma \in R_+^K, \gamma_0 \in R_+}{\operatorname{argmin}} \left(\left\| x - \left(\sum_{j=1}^{K} \gamma_j P_j^{(a)} + \gamma_0 I_P \right) \right\|^2 + \lambda |\gamma|_1 + \beta |\gamma|_2 \right) \qquad (3ter)$$

$$(\hat{\gamma}(\lambda, \beta), \hat{\gamma}_0(\lambda, \beta)) = \underset{\gamma \in R_+^K, \gamma_0 \in R_+}{\operatorname{argmin}} \left(\left\| x - \left(\sum_{j=1}^{K} \gamma_j P_j^{(a)} + \gamma_0 I_P \right) \right\|^2 + \lambda |w_1 \odot \gamma|_1 + \beta |w_2 \odot \gamma|_2 \right) \qquad (3q)$$

expresiones en las que:

- β es el parámetro real positivo;
- $|\ |_2$ es la norma L2; y
- w_1 y w_2 son unos vectores de peso predeterminados de R_+^K.

Según una variante, la selección de las estructuras $\tilde{\gamma}$ de los vectores $\hat{\gamma}$ se realiza mediante un algoritmo de tipo *"stepwise"* simple o estructurado, tal como por ejemplo el algoritmo descrito en el documento "Structured, sparse regression with application to HIV drug resistance" de Daniel Percival *et al.* Annals of Applied Statistics 2011, Vol. 5, Nº 2A, 628-644, o de un enfoque exhaustivo que consiste en ensayar un número importante, incluso la totalidad, de las estructuras posibles para el vector $\hat{\gamma}$.

Según una variante, se omite la etapa 30 de cálculo de los modelos candidatos, siendo los modelos candidatos los obtenidos durante la etapa 12, siendo esta etapa de selección entonces una etapa de cálculo de los modelos candidatos por el algoritmo LASSO.

Asimismo, se han descrito unos modos de realización en los que los microorganismos se referencian a nivel especie.

En una variante, se utilizan varios niveles taxonómicos diferentes, por ejemplo al menos dos niveles entre la especie, la sub-especie y el género.

En una variante, se utilizan otros tipos de caracterización de los microorganismos, en particular unos fenotipos clínicos, como por ejemplo el gram de las bacterias.

Asimismo, se han descrito unos modos de realización aplicados a la espectrometría MALDI-TOF. Son posibles otros tipos de medición, aplicándose la invención a la espectrometría de masa, en particular la espectrometría MALDI-TOF y la espectrometría ESI-MS, a la espectroscopia vibracional, en particular la espectroscopia RAMAN, a la espectroscopia por fluorescencia, en particular la espectroscopia por fluorescencia intrínseca, y a la espectroscopia infrarroja.

Se describirán ahora unos resultados de análisis de muestras biológicas obtenidos según la invención. Más particularmente, se considera una aplicación a la espectrometría MALDI-TOF. Los microorganismos se referencian a nivel especie, los modelos candidatos toman la forma de la relación (1bis), los coeficientes a_{ij} son los coeficientes de Jaccard de la relación (2), la selección de las estructuras $\tilde{\gamma}$ de los vectores $\hat{\gamma}$ se realiza mediante el algoritmo LASSO de la relación (3) cogiendo $\gamma_0 = \hat{\gamma}_0 = 0$, el cálculo de los modelos candidatos se realiza mediante unas relaciones (4bis), (5) y (6) con un $\hat{\gamma}_0$ no forzado a 0, y la selección del modelo candidato $\hat{\gamma}_{sel}$ se realiza según las relaciones (7), (8) y (9).

Se considera un conjunto de K = 20 especies de bacteria de referencia y_j, siendo algunas gram positivas y otras gram negativas, que pertenecen a 9 géneros diferentes, seleccionándose algunas especies debido a la dificultad de distinguirlas por espectrometría de masa. Para cada especie, 11 a 60 espectros de masa se han medido a partir de 7 a 20 cepas de la especie. Un conjunto de 571 espectros de masa para 213 está así constituido.

El vector de intensidad de referencia P_j de cada especie y_j se obtiene aplicando una cuantificación constante entre 3000 y 7000 Thomson con un número p = 1300 intervalos, y para cada intervalo, una intensidad de pico se calcula como se ha descrito anteriormente en la etapa 18 para obtener el vector P_j.

Se han creado unas muestras biológicas mezclando con diferentes ratios dos especies de referencias diferentes, en particular:

- 4 conjuntos de muestras biológicas, referenciadas “A”, “B”, “C”, y “D”, que comprenden dos especies que pertenecen al mismo género;

- 4 conjuntos de muestras biológicas, referenciadas “E” y “F”, que comprenden dos especies que pertenecen a unos géneros diferentes pero que tienen el mismo tipo de gram;

- 4 conjuntos de muestras biológicas, referenciadas “G”, “H”, “I” y “J”, que comprenden dos especies que presentan unos gram diferentes.

Más particularmente, para cada especie de referencia que entra en la constitución de una mezcla, se seleccionan en primer lugar dos cepas diferentes de la especie después, para cada cepa, se produce una muestra “pura” que comprende sólo la cepa. Para obtener un conjunto de muestras biológicas mezclando dos especies, se mezclan después dos pares de muestras puras de las dos especies con los ratios 1:0, 10:1, 5:1, 2:1, 1:1, 1:2, 1:5, 1:10, 0:1.

Se miden después dos espectros de masa y se digitalizan para cada muestra biológica producida, conduciendo a una totalidad de 360 espectros de los cuales 80 corresponden a unas muestras puras. Cada espectro de masa se trata para obtener un vector de intensidad x aplicando la cuantificación utilizada para la construcción de los vectores de intensidad de referencia, y reteniendo el pico de intensidad máximo para cada intervalo de cuantificación.

Las figuras 2A y 2B ofrecen una ilustración del conjunto de datos de test así producido. La figura 2A es una matriz de similitud de los vectores de intensidad de referencia P_j, siendo los coeficientes de la matriz de similitud unos coeficientes de Jaccard. Cuanto más oscuro sea el componente de la matriz de similitud, más importantes será la correlación entre las especies correspondientes. Los cuadrados centrales G1-G9 corresponden a los 9 géneros considerados, el cuadrado G+ a las bacterias gram positivas y el cuadrado G- a las bacterias gram negativas. Los conjuntos de muestra A a J se posicionan después sobre la matriz de similitud. La figura 2B ilustra, por su parte, los picos de los espectros del conjunto D, a saber una mezcla de dos especies que pertenecen al mismo género, los picos de los espectros del conjunto E, a saber una mezcla de dos especies de bacteria que tienen el mismo gram, y los picos de los espectros del conjunto I, a saber una mezcla de dos especies de bacterias de gram diferentes. Cada conjunto ilustrado comprende nueve espectros que corresponden a los diferentes ratios descritos anteriormente, desde el ratio 1:0 que corresponde a una muestra pura de la primera especie al ratio 0:1 que corresponde a una muestra pura de la segunda especie. Los espectros ilustran en particular los picos únicamente presentes en el vector de referencia en la primera especie (picos “Pic1”), los picos únicamente presentes en el vector de referencia de la segunda especie (picos “Pic2”), los picos al mismo tiempo presentes en la primera y en la segunda especies (picos “Pic12”), y los picos que no están presentes ni en la primera especie, ni en la segunda especie (picos « $\overline{\text{Pic12}}$ »). Se señala en particular que para una mezcla de dos especies de bacteria que tienen el mismo género, la gran mayoría de los picos están presentes al mismo tiempo en los dos vectores de referencia de las dos especies, lo que significa que es difícil diferenciarlos. Se observa también que para las mezclas de dos especies de bacteria que tienen el mismo gram, la proporción de picos de una especie presente en el espectro de la mezcla evoluciona de manera coherente con el ratio de esta especie, lo que es menos cierto para las mezclas de dos especies de bacteria que tienen unos gram diferentes.

La capacidad del procedimiento según la invención para detectar una mezcla polimicrobiana y para identificar sus componentes se ha evaluado mediante un criterio de sensibilidad (“sensitivity”) y de un criterio de especificidad (“specificity”) del procedimiento, es decir respectivamente la capacidad del procedimiento para detectar una mezcla de dos especies y una mezcla “pura”. Además, se evalúan también los criterios siguientes: a) la detección de una mezcla microbiana se considera como conseguida cuando se detectan dos o más componentes; b) una mezcla se considera como correctamente identificada cuando se identifican las dos especies que componen la mezcla, y sólo ellas; c) una mezcla se considera como parcialmente identificada cuando se identifica una de las dos especies que componen la mezcla; d) la identificación de una mezcla se considera como un fracaso cuando se identifica una especie que no pertenecen a la mezcla.

La figura 3A ilustra los resultados obtenidos a nivel especie. En términos de detección (gráfico de la izquierda), se observa que se ha detectado el 53,6% de las mezclas polimicrobianas. Por el contrario, se han detectado el 91,2% de las mezclas denominadas “puras”, es decir que comprenden una sola especie, y cerca del 75% de las mezclas polimicrobianas detectadas se han identificado correctamente, elevándose este porcentaje al 86,4% para las mezclas puras. Además, en cuanto a la identificación (gráfico de la derecha), se observa que el 42,1% de las mezclas polimicrobianas han sido perfectamente identificados, llevando a una identificación parcial conseguida en el 82,1% de los casos, y la identificación ha fracasado para aproximadamente el 18% del conjunto de las mezclas polimicrobianas y de las mezclas puras. La gran mayoría de estos fracasos corresponden a unas mezclas que comprenden unas especies del mismo género, lo que es conforme a la dificultad de distinguir unas bacterias que están próximas taxonómicamente.

La conmutación de la detección y la identificación a nivel taxonómico superior, a saber el género, mejora notablemente los resultados como se ilustra en la figura 3B. Los resultados para los géneros se obtuvieron realizando el procedimiento según la invención a nivel especie y después sumando los componentes de los vectores $\hat{\gamma}_{sel}$ para obtener unos resultados a nivel género como se han descrito anteriormente. La sensibilidad y la especificidad de la detección a nivel género alcanzan respectivamente el 61,3% y el 100% para las mezclas polimicrobianas y las mezclas puras, y todos los géneros se identifican de manera sustancial correctamente. Además, se identifican parcialmente las raras mezclas que no se han identificado correctamente. En su totalidad, el 81,4% de las mezclas se han identificado correctamente, habiendo la identificación fracasado para sólo el 0,6% de ellos.

REIVINDICACIONES

1. Procedimiento de detección en una muestra biológica de al menos dos microorganismos que pertenecen a dos taxones diferentes entre un conjunto predeterminado { y_j} de un número de K taxones de referencia y_j diferentes, estando cada taxón de referencia y_j representado por un vector de intensidad de referencia predeterminada P_j de un espacio R^p obtenido sometiendo al menos una muestra biológica de referencia, que comprende un microorganismo que presenta el taxón de referencia, a una técnica de medición que produce una señal numérica multidimensional representativa de la muestra de referencia y determinando dicho vector de referencia en función de dicha señal numérica multidimensional, en la que p es superior a 1, comprendiendo el procedimiento:

▪ la adquisición de una señal numérica multidimensional de la muestra biológica mediante la tecnología de medición;

▪ la determinación de un vector de intensidad x de R^p en función de la señal numérica multidimensional adquirida;

▪ la construcción de un conjunto {$\hat{\gamma}_l$} de modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, \hat{\gamma}_0)_l$ que modelizan el vector de intensidad x según la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} + \hat{\gamma}_0 I_P$$

expresión en la que:

◦ $\hat{x}_l$ es un vector de R^p que reconstruye el vector de intensidad x por el modelo $\hat{\gamma}_l$,

◦ $\hat{\gamma}_0$ escalar es un verdadero y I_P es el vector unidad de R^p;

◦ $\forall j \in [[1,K]]$, $\hat{\gamma}_j$ es el j^{enésimo} componente de un vector $\hat{\gamma}$ de R_+^K;

◦ $\forall j \in [[1,K]]$, $P_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$; y

◦ $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente predeterminado;

▪ la selección de un modelo candidato $\hat{\gamma}_{sel}$ entre el conjunto {$\hat{\gamma}_l$} de los modelos candidatos $\hat{\gamma}_l$, solución de un problema según la relación:

$$\hat{\gamma}_{sel} = \underset{\hat{\gamma}_l \in \{\hat{\gamma}_l\}}{argmin}\left(C_v(\hat{\gamma}_l) + C_c(\hat{\gamma}_l)\right)$$

expresión en la que:

◦ $C_v(\hat{\gamma}_l)$ es un criterio que cuantifica un error de reconstrucción entre el vector de intensidad de la muestra biológica x y la reconstrucción $\hat{x}_l$ del vector de intensidad x por un modelo candidato $\hat{\gamma}_l$; y

◦ $C_c(\hat{\gamma}_l)$ es un criterio que cuantifica la complejidad de un modelo candidato $\hat{\gamma}_l$;

▪ y la determinación de la presencia en la muestra biológica de al menos dos microorganismos que pertenecen a unos taxones diferentes del conjunto predeterminado { y_j} de taxones cuando al menos dos componentes $\hat{\gamma}_j$ del vector $\hat{\gamma}$ del modelo candidato seleccionado $\hat{\gamma}_{sel}$ son superiores a un valor límite predeterminado estrictamente positivo.

2. Procedimiento de identificación de microorganismos presentes en una muestra biológica entre un conjunto predeterminado { y_j} de un número de *K* de taxones de referencia diferentes, estando cada taxón de referencia y_i representado por un vector de intensidad de referencia predeterminada P_j de un espacio R^p obtenido sometiendo al menos una muestra biológica de referencia que comprende un microorganismo que presenta el taxón de referencia a una técnica de medición que produce una señal numérica multidimensional representativa de la muestra de referencia, y determinando dicho vector de referencia en función de dicha señal numérica multidimensional, en la que *p* es superior a 1, comprendiendo el procedimiento:

▪ la adquisición de una señal numérica multidimensional de la muestra biológica mediante la tecnología de medición;

▪ la determinación de un vector de intensidad *x* de R^p en función de la señal numérica multidimensional adquirida;

▪ la construcción de un conjunto {$\hat{\gamma}_l$} de modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, \hat{\gamma}_0)_l$ que modelizan el vector de intensidad *x* según la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} + \hat{\gamma}_0 I_P$$

expresión en la que:

◦ $\hat{x}_l$ es un vector de R^p que reconstruye el vector de intensidad *x* por el modelo $\hat{\gamma}_l$;

◦ $\hat{\gamma}_0$ es un escalar verdadero y I_P es el vector unidad de R^p;

◦ $\forall j \in [[1,K]]$, es el j^{enésimo} componente de un vector $\hat{\gamma}$ de R_+^K;

◦ $\forall j \in [[1,K]]$, $\hat{P}_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$; y

◦ $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente predeterminado;

▪ la selección de un modelo candidato $\hat{x}_{sel}$ entre el conjunto {$\hat{x}_l$} de los modelos candidatos $\hat{x}_l$, solución de un problema según la relación:

$$\hat{x}_{sel} = \underset{\hat{x}_l \in \{\hat{x}_l\}}{argmin}\left(C_v(\hat{x}_l) + C_c(\hat{x}_l)\right)$$

expresión en la que:

◦ $C_v(\hat{x}_l)$ es un criterio que cuantifica un error de reconstrucción entre el vector de intensidad de la muestra biológica *x* y un modelo candidato $\hat{x}_l$; y

◦ $C_c(\hat{x}_l)$ es un criterio que cuantifica la complejidad de un modelo candidato $\hat{x}_l$;

▪ y la determinación de la presencia en la muestra biológica de un microorganismo de taxón y_j del conjunto predeterminado { y_j} para cada componente $\hat{\gamma}_j$ del vector $\hat{\gamma}$ del modelo candidato seleccionado superior a un valor límite predeterminado estrictamente positivo.

3. Procedimiento de determinación de la abundancia relativa en una muestra biológica de al menos dos microorganismos que pertenecen a dos taxones diferentes entre un conjunto predeterminado {y_j} de un número de *K* taxones de referencia y_j diferentes, estando cada taxón de referencia y_j representado por un vector de intensidad de referencia predeterminada P_j de un espacio R^p obtenido sometiendo al menos una muestra biológica de referencia que comprende un microorganismo que presenta el taxón de referencia a una técnica de medición que produce una señal numérica multidimensional representativa de la muestra de referencia y determinando dicho vector de referencia en función de dicha señal numérica multidimensional, en la que *p* es superior a 1, comprendiendo el procedimiento:

▪ la adquisición de una señal numérica multidimensional de la muestra biológica mediante la tecnología de medición;

▪ la determinación de un vector de intensidad *x* de R^p en función de la señal numérica multidimensional adquirida;

▪ la construcción de un conjunto $\{\hat{\gamma}_l\}$ de modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, \hat{\gamma}_0)_l$ que modelizan el vector de intensidad *x* según la relación:

$$\hat{x}_l = \sum_{j=1}^{K} \hat{\gamma}_j P_j^{(a)} + \hat{\gamma}_0 I_P$$

expresión en la que:

◦ $\hat{x}_l$ es un vector de R^p que reconstruye el vector de intensidad *x* por el modelo $\hat{\gamma}_l$;

◦ $\hat{\gamma}_0$ escalar es un verdadero y I_P es el vector unidad de R^p;

◦ $\forall j \in [[1,K]]$, $\hat{\gamma}_j$ es el j^enésimo componente de un vector $\hat{\gamma}$ de R_+^K;

◦ $\forall j \in [[1,K]]$, $P_j^{(a)} = \sum_{i=1}^{K} a_{ij} P_i$; y

◦ $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente predeterminado;

▪ la selección de un modelo candidato $\hat{\gamma}_{sel}$ entre el conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l$, solución de un problema según la relación:

$$\hat{\gamma}_{sel} = \underset{\hat{\gamma}_l \in \{\hat{\gamma}_l\}}{argmin}\left(C_v(\hat{\gamma}_l) + C_c(\hat{\gamma}_l)\right)$$

expresión en la que:

◦ $C_v(\hat{\gamma}_l)$ es un criterio que cuantifica un error de reconstrucción entre el vector de intensidad de la muestra biológica *x* y la reconstrucción $\hat{x}_l$ del vector de intensidad *x* por un modelo candidato $\hat{\gamma}_l$; y

◦ $C_c(\hat{\gamma}_l)$ es un criterio que cuantifica la complejidad de un modelo candidato $\hat{\gamma}_l$;

▪ y la determinación de la abundancia relativa en la muestra biológica relativa C_j de un taxón de referencia y_j según la relación:

$$C = J(\hat{\gamma}_{sel})$$

expresión en la que *J* es una función matricial de $R_+^p \times R_+^K$ en R_+^K y $C = (C_1 \ldots C_j \ldots C_K)^T$ es un vector de R_+^K con $\forall j \in [[1,K]]$, C_j es la abundancia relativa du taxón de referencia y_j.

4. Procedimiento según la reivindicación 1, 2 o 3, en el que $\forall (i,j) \in [[1,K]]^2$, a_{ij} es un coeficiente de similitud entre los vectores de referencia P_i y P_j de los taxones de referencia y_i y y_j.

5. Procedimiento según la reivindicación 4, en el que el coeficiente de similitud a_{ij} entre los vectores de referencia P_i y P_j es igual al coeficiente de Jaccard entre unas versiones binarizadas de los vectores P_i y P_j.

6. Procedimiento según una cualquiera de las reivindicaciones anteriores, en el que $\hat{\gamma}_0 = 0$, y en el que la construcción del conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, 0)_l$ comprende la resolución de un conjunto de

problemas de optimización para unos valores de un parámetro λ de R_+, estando cada problema definido según la relación:

$$\hat{\gamma}(\lambda) = \underset{\gamma \in R_+^K}{\operatorname{argmin}} \left(\left\| x - \sum_{j=1}^{K} \gamma_j P_j^{(a)} \right\|^2 + \lambda |\gamma|_1 \right)$$

expresión en la que $|\gamma|_1$, es la norma L1 del vector γ.

7. Procedimiento según una cualquiera de las reivindicaciones 1 a 5, en el que $\hat{\gamma}_0 = 0$, y en el que la construcción del conjunto $\{\hat{\gamma}_l\}$ de los modelos candidatos $\hat{\gamma}_l = (\hat{\gamma}, 0)_l$ comprende la resolución de un conjunto de problemas de optimización para unos valores de parámetros λ y β de R_+, estando cada problema definido según la relación:

$$\hat{\gamma}(\lambda, \beta) = \underset{\gamma \in R_+^K}{\operatorname{argmin}} \left(\left\| x - \sum_{j=1}^{K} \gamma_j P_j^{(a)} \right\|^2 + \lambda \left| w_1 \odot \gamma \right|_1 + \beta \left| w_2 \odot \gamma \right|_2 \right)$$

expresión en la que:

- $|\ |_1$ es la norma L1;
- $|\ |_2$ es la norma L2;
- $a \odot b$ es el producto término por término de los vectores a y b; y
- w_1 y w_2 son unos vectores de pesos predeterminados de R_+^K.

8. Procedimiento según una de las reivindicaciones 6 o 7, en el que para cada vector $\hat{\gamma}$ solución de un problema de optimización, se calcula un nuevo modelo candidato $\hat{\gamma}_l = \left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right)_l$, y se sustituye el modelo $\hat{\gamma}_l = (\hat{\gamma}, 0)_l$ que corresponde al vector $\hat{\gamma}$, estando los componentes del vector $\hat{\gamma}^{lm}$ del nuevo modelo $\hat{\gamma}_l = \left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right)_l$, que corresponde a los componentes nulos del vector $\hat{\gamma}$, forzados a cero, y calculándose el nuevo modelo $\hat{\gamma}_l = \left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right)_l$ resolviendo el problema de optimización según las relaciones:

$$\left(\hat{\gamma}^{lm}, \hat{\gamma}_0^{lm}\right) = \underset{\substack{\gamma_0^{lm} \in R_+ \\ \gamma^{lm} \in R_+^K}}{\operatorname{argmax}} \left(-\frac{p}{2} ln(2\pi\sigma(x_l)^2) - \frac{1}{2\sigma(x_l)^2} \sum_{b=1}^{p} (x_b - x_{lb})^2 \right)$$

$$\sigma(x_l)^2 = \frac{1}{p} \sum_{b=1}^{p} (x_b - x_{lb})^2$$

$$x_l = \gamma_0^{lm} I_p + \sum_{j:\hat{\gamma}_j \neq 0} \gamma_j^{lm} P_j^{(a)}$$

expresiones en las que:

- x_b es el b^{enésimo} componente del vector de intensidad de la muestra biológica x; y
- x_{lb} es el b^{enésimo} componente del vector de reconstrucción $x_l = \gamma_0^{lm} I_p + \sum_{j:\hat{\gamma}_j > 0} \gamma_j^{lm} P_j^{(a)}$.

9. Procedimiento según una cualquiera de las reivindicaciones anteriores, en el que el criterio $C_v(\hat{\gamma}_l)$ que cuantifica el error de reconstrucción es un criterio de verosimilitud.

10. Procedimiento según la reivindicación 9, en el que:

$$C_v(\hat{\gamma}_l) = -\frac{p}{2} ln(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{b=1}^{p}(x_b - \hat{x}_{lb})^2$$

expresión en la que:

- $\hat{\sigma}^2 = \frac{1}{p}\sum_{b=1}^{p}(x_b - \hat{x}_{lb})^2$;
- x_b es el b^{enésimo} componente del vector de picos de la muestra biológica x; y
- $\hat{x}_{lb}$ es el b^{enésimo} componente del vector de reconstrucción $\hat{x}_l$ del modelo candidato $\hat{\gamma}_l$.

11. Procedimiento según una cualquiera de las reivindicaciones anteriores, en el que el criterio $C_c(\hat{\gamma}_l)$ que cuantifica la complejidad del modelo $\hat{\gamma}_l$ cuantifica dicha complejidad en términos de número de componentes $\hat{\gamma}_j$ del vector $\hat{\gamma}$ estrictamente positivos.

12. Procedimiento según la reivindicación 11, en el que:

$$\text{Si } \hat{\gamma}_0 = 0 \quad \text{entonces} \quad C_c(\hat{\gamma}_l) = \left(1 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) \ln p$$

$$\text{Si } \hat{\gamma}_0 \neq 0 \quad \text{entonces} \quad C_c(\hat{\gamma}_l) = \left(2 + \sum_{j=1}^{K} \mathbf{1}\left(\hat{\gamma}_j > 0\right)\right) \ln p$$

expresiones en las que la función 1(.) es igual a 1 si su argumento es verdadero y cero si no lo es.

13. Procedimiento según una cualquiera de las reivindicaciones anteriores, en el que los taxones pertenecen a un mismo nivel taxonómico, en particular el nivel especie, género o sub-especie.

14. Procedimiento según una cualquiera de las reivindicaciones 1 a 12, en el que los taxones pertenecen a al menos dos niveles taxonómicos diferentes, en particular especies, géneros, y/o sub-especies.

15. Procedimiento según una cualquiera de las reivindicaciones anteriores, en el que los taxones pertenecen a un primer nivel taxonómico, y en el que un modelo del vector x se calcula para un segundo nivel taxonómico superior al primer nivel taxonómico sumando los componentes del vector $\hat{\gamma}$ que corresponde a los taxones que dependen de dicho nivel taxonómico superior.

16. Procedimiento según la reivindicación 15, en el que el modelo del vector x se calcula para el nivel taxonómico superior si un grado de similitud dentro del primer nivel es superior a un límite predeterminado.

17. Procedimiento según una cualquiera de las reivindicaciones 1 a 14, en el que si un grado de similitud entre un conjunto de taxones definido dentro de un primer nivel es superior a un límite predeterminado, entonces para la constitución del conjunto predeterminado $\{ y_j \}$ unos taxones de referencia, denominados taxones, son reagrupados y sustituidos por un taxón de referencia definido a un segundo nivel taxonómico, superior al primer nivel taxonómico.

18. Procedimiento según una cualquiera de las reivindicaciones anteriores, en el que la técnica de medición produce un espectro y en el que los vectores de intensidad de referencia P_j son unas listas de picos comprendidos en los espectros de los taxones de referencia y_j.

19. Procedimiento según la reivindicación 18, en el que la técnica de medición comprende una espectrometría de masa.

20. Procedimiento según una cualquiera de las reivindicaciones 3 a 19, en el que:

$$C_j = \frac{\hat{\gamma}_{j,sel}}{\sum_{i=1}^{K} \hat{\gamma}_{i,sel}}$$

expresión en la que $\forall j \in [[1,K]]$, $\hat{\gamma}_{j,sel}$ es el j^{enésimo} componente del vector $\hat{\gamma}$ del modelo seleccionado $\hat{\gamma}_{sel}$.

21. Dispositivo de análisis de una muestra biológica que comprende:

- un espectrómetro o un espectroscopio apto para producir unos espectros de la muestra biológica;
- una unidad de cálculo apta para realizar un procedimiento según una cualquiera de las reivindicaciones anteriores.

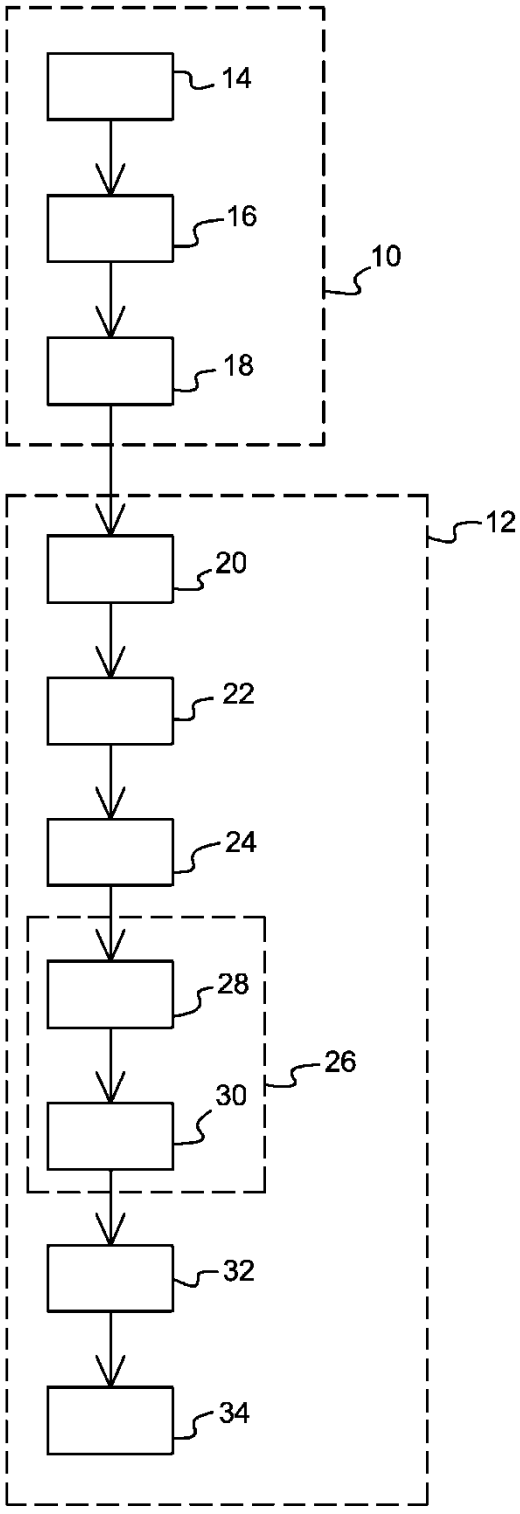

Fig. 1

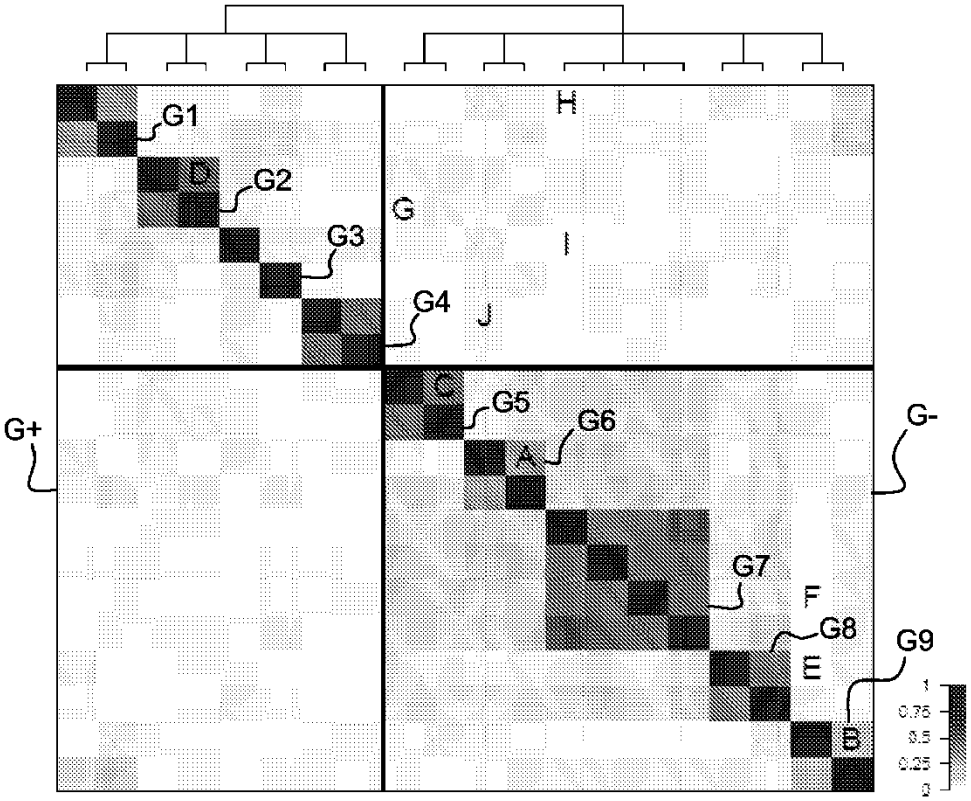

Fig. 2A

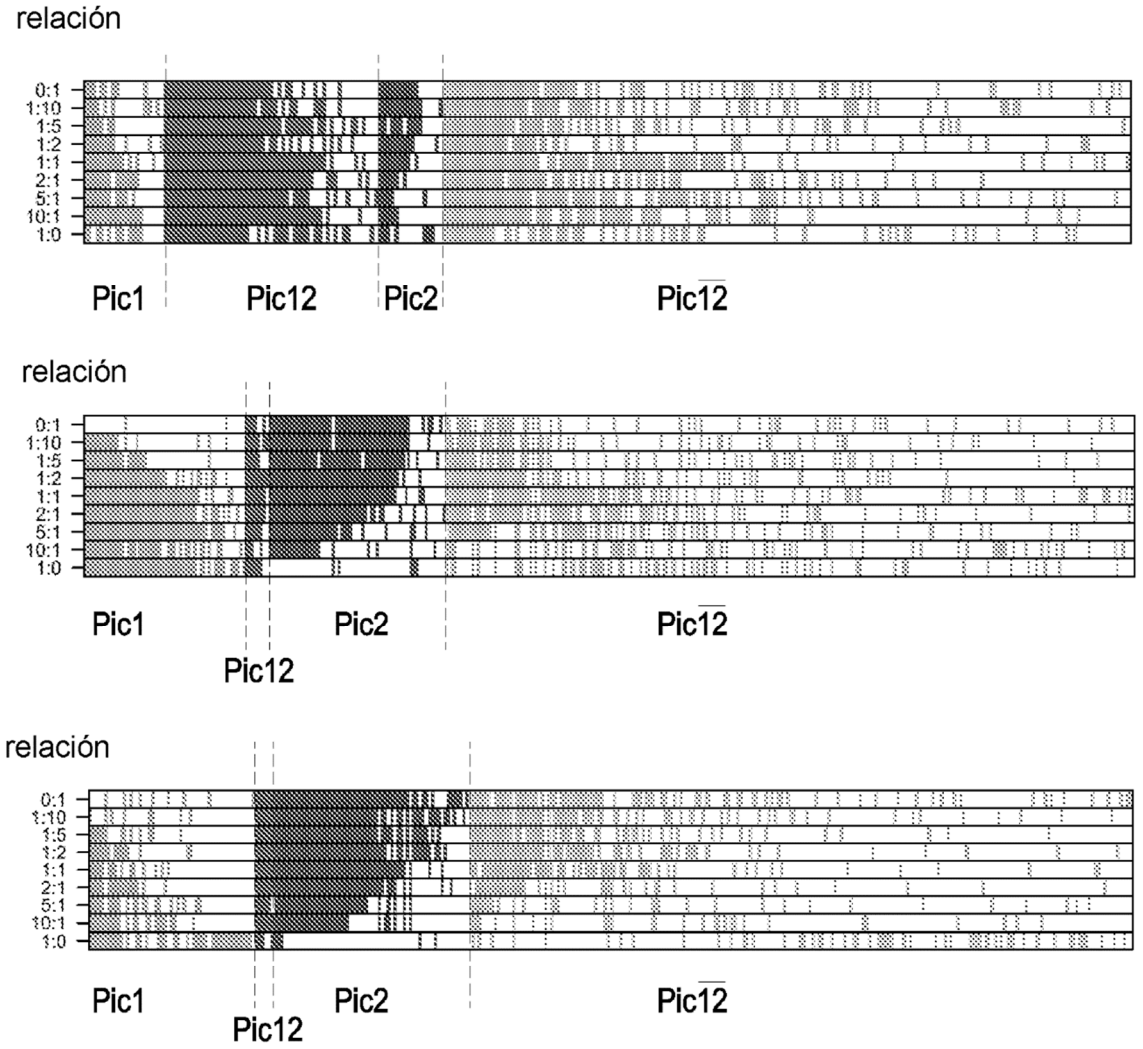
relación
0:1
1:10
1:5
1:2
1:1
2:1
5:1
10:1
1:0
Pic1
Pic12
Pic2
Pic12
relación
Pic1
Pic2
Pic12
Pic12
relación
Pic1
Pic12
Pic2
Pic12
picos presentes en el prototipo de la 1ª especie
picos presentes en el prototipo de la 2ª especie
picos presentes en ambos prototipos
picos ausentes en ambos prototipos

Fig. 2B

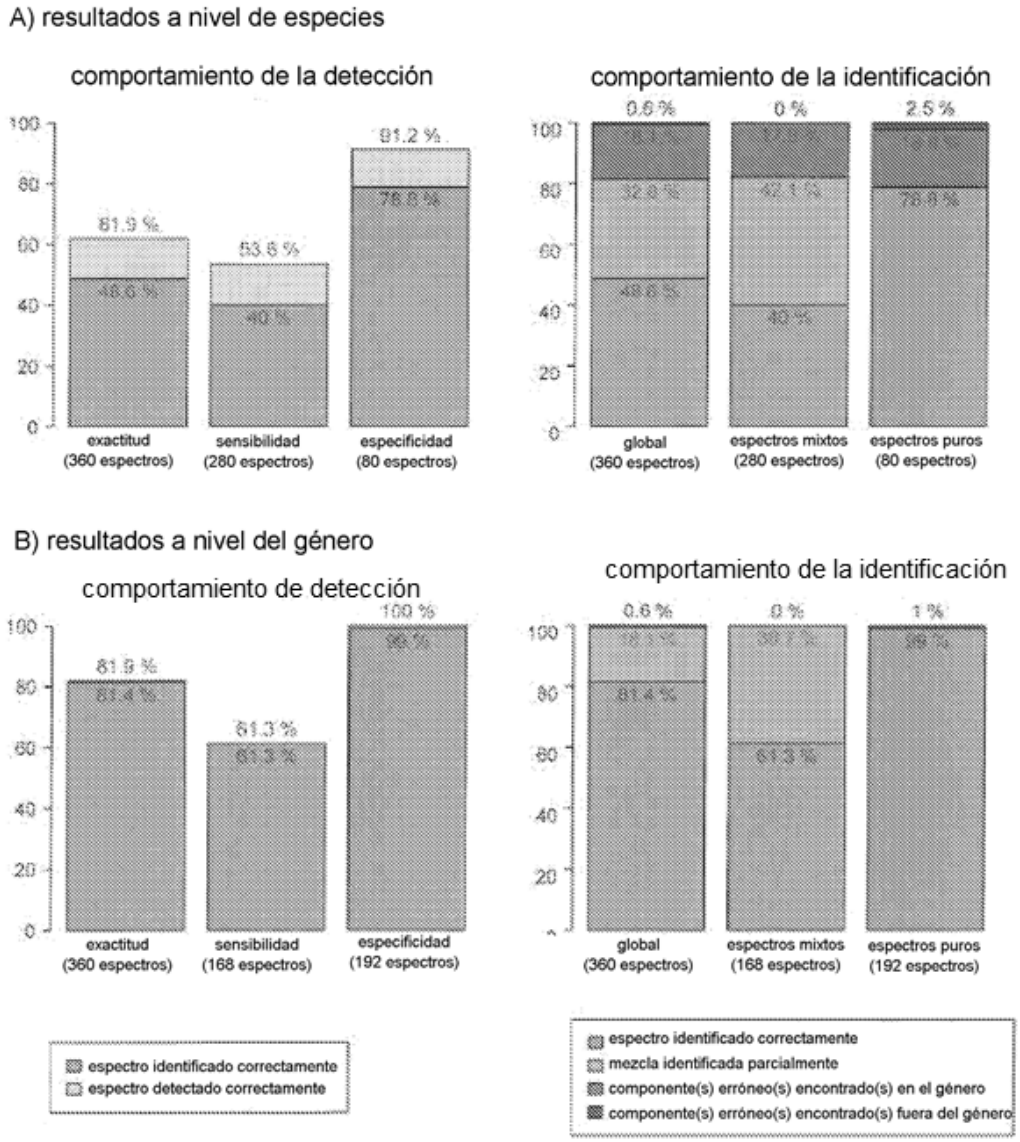
A) resultados a nivel de especies
comportamiento de la detección
100
80
60
40
20
0
81.9 %
48.6 %
53.6 %
40 %
81.2 %
78.8 %
exactitud (360 espectros)
sensibilidad (280 espectros)
especificidad (80 espectros)
comportamiento de la identificación
0.6 %
32.8 %
48.6 %
0 %
42.1 %
40 %
2.5 %
78.8 %
global (360 espectros)
espectros mixtos (280 espectros)
espectros puros (80 espectros)
B) resultados a nivel del género
comportamiento de detección
81.9 %
81.4 %
61.3 %
61.3 %
100 %
96.5 %
exactitud (360 espectros)
sensibilidad (168 espectros)
especificidad (192 espectros)
comportamiento de la identificación
0.6 %
81.4 %
0 %
61.3 %
1 %
96.5 %
global (360 espectros)
espectros mixtos (168 espectros)
espectros puros (192 espectros)
espectro identificado correctamente
espectro detectado correctamente
espectro identificado correctamente
mezcla identificada parcialmente
componente(s) erróneo(s) encontrado(s) en el género
componente(s) erróneo(s) encontrado(s) fuera del género

Fig. 3