(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 432 677**

(51) Int. CI.:

G06F 19/18 (2011.01)
G01N 33/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **30.06.2010 E 10793643 (7)**

(97) Fecha y número de publicación de la concesión europea: **07.08.2013 EP 2450815**

(54) Título: **Método de identificación de péptidos y proteínas a partir de datos de espectrometría de masas**

(30) Prioridad:

01.07.2009 ES 200930402

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
04.12.2013

(73) Titular/es:

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (100.0%)
C/ Serrano 117
28006 Madrid, ES

(72) Inventor/es:

ALBAR RAMÍREZ, JUAN PABLO y
RAMOS FERNÁNDEZ, ANTONIO

(74) Agente/Representante:

ILLESCAS TABOADA, Manuel

DESCRIPCIÓN

Método de identificación de péptidos y proteínas a partir de datos de espectrometría de masas

CAMPO DE LA INVENCIÓN

La presente invención se inscribe dentro del campo de los métodos de identificación y caracterización estructural de proteínas a gran escala mediante técnicas de espectrometría de masas.

ANTECEDENTES DE LA INVENCIÓN

La Proteómica es una de las ciencias de la era post-genómica que posee un mayor impacto en la biotecnología moderna, pues comprende la identificación y cuantificación de grandes cantidades de proteínas en matrices extremadamente complejas (fluidos biológicos, tejidos o cultivos celulares, entre otras). Actualmente, las técnicas de mayor éxito y relevancia académica e industrial empleadas en proteómica son aquéllas basadas en espectrometría de masas en tándem (MS/MS), que consisten en la extracción de las proteínas de la muestra a analizar, la digestión de dichas proteínas con enzimas u otros agentes químicos para obtener péptidos (más fáciles de analizar), separar dichos péptidos habitualmente mediante técnicas cromatográficas, e introducirlos en un espectrómetro de masas en forma ionizada para medir su masa y fragmentarlos dentro del espectrómetro con el objetivo de obtener información estructural, de modo que permita la identificación de las proteínas conformadas por los péptidos analizados.

La investigación actual en Proteómica basada en espectrometría de masas en tándem comprende la generación de grandes volúmenes de datos que contienen típicamente entre miles y millones de espectros de masas. Dichos espectros son asignados a secuencias de péptidos registradas en bases de datos, empleando programas informáticos denominados motores de búsqueda. En el desarrollo histórico de la Proteómica basada en MS/MS, dado el alto número de espectros involucrados en los análisis, la validación manual de la correspondencia espectro-péptido se ha convertido en poco tiempo en impracticable, por lo que se ha hecho necesario el desarrollo de procedimientos automáticos no manejados por el usuario, que permitan identificar los péptidos analizados, así como descartar las correspondencias espurias (conocidas como falsas detecciones o falsos positivos). Estos procedimientos comprenden el empleo de algoritmos basados en sistemas de puntuación estadística para clasificar cada espectro analizado en una muestra, de forma que, cuanto mayor sea la puntuación obtenida, mayor es la probabilidad de que la identificación espectro-péptido sea la correcta.

Actualmente, las diferencias existentes entre los distintos motores de búsqueda del mercado se derivan del pre-procesado y la normalización de los espectros MS/MS analizados, como consecuencia del empleo de distintos modelos estadísticos y métodos numéricos en el sistema de puntuación de cada motor. Estas diferencias suponen el principal problema a la hora de analizar espectros MS/MS empleando múltiples motores de búsqueda, ya que algunas secuencias de péptidos identificadas correctamente en alguno de los motores, pueden no serlo en otros. Éste es un hecho ampliamente conocido por los espectrometristas experimentados. La presente invención comprende un método de búsqueda combinada empleando múltiples motores (definida de aquí en adelante como meta-búsqueda) orientado a la solución de este inconveniente, así como a la optimización de las técnicas de análisis de los espectros obtenidos mediante MS/MS. Este método proporciona también un criterio generalizado de puntuación (que definimos como meta-puntuación) de los resultados obtenidos por los distintos motores de bases de datos empleados, mediante una modelización estadística suficientemente robusta que permita obtener una identificación espectro-péptido única.

A pesar de los beneficios potenciales que posee un método de meta-búsqueda con múltiples motores, pocos son hasta la fecha los intentos que se han realizado en esta dirección. Entre los más relevantes, cabe citar los trabajos desarrollados por Rohrbough *et al* [1], Higgs *et al* [2], Searle *et al* [3] y Alves *et al* [4]. Por otra parte, dentro de estado de la técnica relacionado con la investigación en proteómica, es más abundante la existencia de productos comerciales con opciones de búsqueda comparativa (lo que difiere del concepto de meta-búsqueda) utilizando varios motores que presentan algunas aplicaciones informáticas del mercado, tales como la opción “InChorus” del motor de búsqueda PEAKS (distribuido por Bioinformatics Solutions Inc.), el sistema de análisis de datos Rosetta Elucidator (distribuido por Rosetta Biosoftware), la plataforma de análisis Proteome Discoverer (distribuida por Thermo Fisher Scientific Inc.) o el motor Phenyx, distribuido por Geneva Bioinformatics SA.

Otra aplicación de este campo de la técnica es la implementación de los métodos de búsqueda en dispositivos de análisis de péptidos y proteínas que combinan tanto hardware como software, y son comercializados de forma autónoma como estaciones de trabajo “plug-and-play” o como servidores que permiten ser empleados simultáneamente por múltiples usuarios. Un ejemplo de este tipo de dispositivos sería la estación de trabajo Sorcerer 2, comercializada por la empresa Sage-N Research, Inc., o el servidor configurable distribuido de forma conjunta por IBM y Thermo Electron Corporation. Estos dispositivos tampoco integran, hasta la fecha, el uso simultáneo de varios motores mediante un método de meta-búsqueda.

Si bien la presente invención comparte algunos planteamientos y objetivos con cada una de las técnicas anteriormente citadas, es el único de todos los métodos que presenta el siguiente conjunto de ventajas:

- El método de meta-búsqueda y su sistema de meta-puntuación agrega información adicional que no puede ser obtenida mediante la búsqueda con un solo motor.

- Emplea una modelización estadística robusta que permite la elección de una única combinación de secuencia de péptidos, carga eléctrica y composición química por espectro (a diferencia de los métodos empleados por PEAKS, Rosetta Elucidator, Proteome Discoverer y Phenyx, que únicamente usan los resultados de múltiples motores con fines comparativos, sin la posibilidad de utilizar una estadística común y un sistema común de meta-puntuación).

- Este método es completamente generalizable para el empleo de cualquier número de motores de búsqueda (a diferencia de los métodos propuestos en las Referencias [1] y [2], cuya generalización a más de dos motores no resulta factible).

- Emplea un método estándar aplicable a los resultados de cualquier motor de búsqueda para obtener las funciones de distribución estadística, a diferencia del método descrito en la Referencia [3] y su implementación comercial en la aplicación Scaffold (distribuida por Proteome Software Inc.), cuya extensión a más de los tres motores estudiados necesitaría encontrar una distribución satisfactoria para cada nuevo motor de búsqueda utilizado.

- Integra en su formulación el empleo de parámetros de concordancia, definidos como el número de otros motores de búsqueda que han proporcionado el mismo péptido candidato que un motor dado. El empleo de parámetros de concordancia no se contempla en el método planteado en la Referencia [4], perdiéndose a causa de su ausencia una parte valiosa de la información, que contribuye sensiblemente al incremento del número péptidos identificados.

- Optimiza automáticamente los valores de todos los parámetros involucrados en el proceso a través de modelado estadístico, sin que sea necesario definir ningún otro tipo de filtro, mecanismo de puntuación arbitraria o predefinir valores para los coeficientes de estos últimos, a diferencia de los métodos basados en mecanismos arbitrarios de filtros múltiples o de puntuación arbitraria descritos en las referencias [4] y [5].

- En cuanto a la detección de proteínas, se emplea un método estadístico riguroso, no sesgado, que emplea un filtrado definido por las tasas de error en las asignaciones secuencia-péptido.

- Adicionalmente, el método reivindicado es suficientemente flexible como para incorporar otras fuentes de información adicionales a la concordancia del motor, tales como el filtrado mediante el error de masa del ión precursor de la secuencia (definido como la diferencia entre la masa teórica de un ión de péptido y la medición de la masa obtenida por el espectrómetro, ya sea utilizando su masa molecular o su relación masa/carga, m/Z), el error en el tiempo de retención (definido como el tiempo característico de retención durante la separación cromatográfica), el error de predicción del punto isoeléctrico (similar al factor anterior, cuando los péptidos se fraccionan utilizando técnicas de separación por isoelectroenfoque), la movilidad iónica (en los espectrómetros de masas que incorporan ese tipo de análisis, basado en la acumulación iónica de especies químicas bajo la acción de un campo eléctrico), la especificidad de la digestión enzimática empleada (es decir, las características de la segmentación de las proteínas en función del tipo de enzimas empleados para su digestión), la detección de múltiples patrones isotópicos para un mismo péptido (habitual en experimentos de marcado isotópico estable empleados en aplicaciones de proteómica cuantitativa) o la concordancia con la secuenciación obtenida por MS/MS sin el uso de un motor de búsqueda (conocida como secuenciación *de novo* de la información). Esta flexibilidad permite al método de meta-búsqueda la integración de datos empleando diferentes preparaciones de muestras, distintos métodos de digestión de proteínas y diversos mecanismos de fragmentación de iones, lo que lo convierten en una herramienta adecuada para la identificación a gran escala de proteínas.

La presente invención se basa en un método de meta-búsqueda empleando los resultados de identificación espectro-péptido obtenidos en diferentes motores de búsqueda sobre bases de datos híbridas diana/señuelo, que contienen una proporción 1:1 de proteínas reales frente a proteínas falsas. Dichas proteínas falsas se obtienen habitualmente invirtiendo la secuencia de cada una de las proteínas reales. Como paso previo a la identificación de meta-resultados, el método de análisis de resultados en cada uno de los motores estudiados por separado se realiza mediante la técnica desarrollada por Ramos-Fernández *et al* [5] (desarrollada para búsquedas empleando un único motor), basada en el uso de distribuciones Lambda generalizadas (GLD's). Dichas GLD's son funciones de cuatro parámetros extremadamente flexibles que pueden representar con gran precisión la mayoría de las familias más importantes de distribuciones de probabilidad continuas empleadas en modelización estadística de histogramas. El modelo de GLD's (descrito en, por ejemplo, el trabajo de Karian *et al* [6]) no ha sido previamente empleado para realizar búsquedas combinadas en múltiples motores de bases de datos de secuencias, y proporciona el marco teórico del modelo estadístico sobre el que opera el método de meta-búsqueda y meta-puntuación aquí reivindicado. A diferencia del modelo de la Referencia [6], la invención aquí reivindicada se presenta como un método que pueda ser implementado de forma automática, suministrando criterios objetivos que permitan la elección de la GLD que mejor se adapte a los resultados observados, sin necesidad de supervisar personalmente cada uno de los modelos candidatos.

REFERENCIAS

[1] Rohrbough, J.G., Brescia, L., Merchant, N., Miller, S., Haynes, P.A. (2006). "Verification of single-peptide protein identifications by the application of complementary database search programs". *J. Biomol. Tech.* 5, 327-332.
[2] Higgs, R.E., Knierman, M.D., Freeman A.B., Gelbert, L.M., Patil, S.T., Hale, J.E. (2007). "Estimating the statistical significance of peptide identifications from shotgun proteomics experiments". *J. Proteome Res.* 6, 1758-1767.
[3] Searle B. C., Turner M., Nesvizhskii A. (2008). "Improving sensitivity by combining results from multiple MS/MS search methodologies". *J. Proteome Res.* 7, 245-253.
[4] Alves, G., Wu, W.W., Wang, G., Shen, R.-F., Yu, Y.-K. (2008). "Enhancing peptide identification confidence by combining search methods". *J. Proteome Res.* 8, 3102-3113.
[5] Yu W., Taylor JA, Davis MT, Bonilla LE, Lee KA, Auger PL, Farnsworth CC, Welcher AA, Paternson SD (2010). "Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by Harnessing multiple engines". *Proteomics 10*, 1172-1189.
[6] Ramos-Fernández, A., Paradela, A., Navajas, R., Albar, J.P. (2008). "Generalized method for probability-based peptide and protein identification from tandem mass spectrometry data and sequence database searching". *Mol. Cell. Proteomics* 7, 1748-1754.
[7] Karian Z.A., Dudewicz, E.J. (2000) "Fitting statistical distributions: the Generalized Lambda Distribution and Generalized Bootstrap methods". Chapman and Hall/CRC.

SUMARIO DE LA INVENCIÓN

Un objeto de la presente invención es proporcionar un método generalizado para la identificación de péptidos y proteínas a partir de datos de espectrometría de masas en tándem.

Otro objeto de la presente invención es proporcionar un motor de meta-búsqueda en el cual los péptidos candidatos se obtienen a partir de múltiples motores de búsqueda en bases de datos de secuencias.

Éstos y otros objetos se consiguen mediante un método en el que:

- Se realiza una búsqueda empleando, al menos, dos motores de búsqueda en bases de datos de secuencias (meta-búsqueda) y que puede ser extendido para el análisis de cualquier número de motores. Esto genera información adicional que no puede ser obtenida mediante la búsqueda con un solo motor.

- Se clasifican los péptidos candidatos en cada motor de búsqueda para construir un modelo de distribuciones Lambda generalizadas (GLD's). Se consigue con ello un soporte teórico completamente general, aplicable a un número arbitrario de motores de búsqueda.

- Se integran los datos de los múltiples motores de búsqueda mediante un sistema de meta-puntuación basado en distribuciones generalizadas de probabilidad y valores-p generalizados (definidos estos últimos como los valores de probabilidad de que una determinada detección secuencia-péptido se haya producido de forma aleatoria). Se consigue con ello una modelización estadística robusta que permite la elección de una única combinación secuencia de péptidos, carga eléctrica y composición química, por espectro.

- El sistema de meta-puntuación incluye la presencia de parámetros de concordancia que proporcionan información sobre la concordancia de asignaciones secuencia-péptido en múltiples motores. Se consigue con ello obtener una mayor información de análisis, que contribuye sensiblemente al incremento del número de péptidos identificados.

En una realización preferente de la presente invención, la integración de los datos de los múltiples motores de búsqueda se lleva a cabo mediante un sistema de meta-puntuación basado en distribuciones Lambda generalizadas (GLD's) y valores-*p* generalizados. Se consigue con ello una distribución única de meta-puntuaciones, así como un sistema de clasificación de identificación secuencia-péptido que integra los datos de todos los motores de búsqueda utilizados, proporcionando información agregada no disponible mediante el uso de un único motor.

En una realización preferente de la presente invención se estima una tasa de error generalizada, bien por medio de la tasa de falsa detección (nombrada habitualmente por su término inglés, "false discovery rate", o FDR, y definida en la descripción detallada de la invención), bien por medio de la tasa de impacto en señuelo (designada por su término inglés, "decoy hit rate" o DHR, y definida en la descripción detallada de la invención), por medio de la probabilidad de obtener al menos un falso positivo (denominada esta tasa como "Familywise error rate", FWER, y definida en la descripción detallada de la invención) o por medio de cualquier otra medida estadística del error en la identificación. Se consigue con ello facilitar un estimador del acierto en las asignaciones péptido-secuencia para un conjunto de datos dado.

En una realización preferente de la presente invención, se calculan los valores-*p* correspondientes a la identificación de proteínas precursoras de los conjuntos de datos, así como las tasas de error FDR y DHR de dichas proteínas. De esta

forma, se consigue con ello un conjunto único de datos sobre la información agregada de todos los motores de búsqueda sobre la proteína precursora de los péptidos detectados.

En una realización preferente de la presente invención, se asigna un coeficiente de peso distinto a cada motor durante la fase de meta-puntuación, siendo éste establecido *a priori* o calculado en función de factores tales como la tendencia a la concordancia de algunos de los motores seleccionados (por emplear, por ejemplo, algoritmos similares que produzcan solapamiento de resultados), que alguno de los motores posea un rendimiento muy superior al resto, o cualquier otra situación en la que se desee efectuar una ponderación asimétrica entre las distintas fuentes empleadas. Con ello se incorpora la posibilidad de favorecer el valor de la información obtenida por unos motores sobre otros.

En una realización preferente de la presente invención, se establece una relación entre la meta-puntuación calculada para una identificación espectro-péptido y las características de la secuencia del péptido candidato, tales como su longitud, presencia o ausencia de sub-secuencias o motivos estructurales, así como la concordancia de la secuencia del péptido con lo esperado a partir del mecanismo de corte del agente químico utilizado en la digestión de las proteínas. Se consigue con ello incorporar al método de meta-puntuación aquellos factores esperables en las secuencias obtenidas, en función de las características de experimento analizado, para mejorar la discriminación entre asignaciones correctas e incorrectas.

En una realización preferente de la presente invención, se integra el método de meta-búsqueda en un dispositivo destinado al análisis de resultados de espectrometría de masas en tándem, que comprenda los medios mecánicos, electromagnéticos, electrónicos o informáticos realizados en forma de hardware y/o software, estando éstos orientados a conformar un sistema de análisis de datos para la identificación de péptidos y proteínas.

Otras características y ventajas de la presente invención se desprenderán de la descripción detallada que sigue y de una realización ilustrativa de su objeto en relación con la figura que lo acompaña.

DESCRIPCIÓN DE LAS FIGURAS

La Figura 1 es un diagrama esquemático del método de meta-búsqueda descrito en la presente invención. En él se representa la búsqueda de secuencias MS/MS sobre conjuntos de espectros **1** mediante el uso de múltiples motores de búsqueda **M** disponibles en el mercado sobre bases de datos híbridas diana/señuelo **2**. Las puntuaciones **x** asociadas a los resultados señuelo se clasifican por el estado de carga del ión precursor y se representan como densidades de probabilidad **y**, ajustándose a un modelo GLD y calculando sus valores-p **V**. Los valores-p **V** obtenidos se representan frente a su frecuencia relativa **x'**. Se incluye también el modelo GLD utilizado para representar la distribución de las meta-puntuaciones **x''** como densidades de probabilidad **y**, realizada durante la fase de cálculo de meta-puntuación **3** descrito por la presente invención.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

El método de meta-búsqueda reivindicado en la presente invención comprende las siguientes etapas:

1. Búsqueda MS/MS: La búsqueda en las bases de datos MS/MS se realiza por medio de los motores de búsqueda empleados en el proceso de meta-búsqueda. La lista de picos de los espectros se emplea como input del sistema, determinando los parámetros de cada motor de búsqueda de acuerdo a un esquema común, fijando la tolerancia de masa precursora y la tolerancia de masa del ión fragmento (es decir, los errores tolerados en los valores de las masas calculadas), la especificidad de la digestión enzimática (es decir, el tipo de fragmentación producida por el enzima empleado para digerir las proteínas), o cualesquiera otros parámetros en función del motor empleado y el conjunto de datos analizado.

2. Ajustes GLD: Se emplea un modelo basado en distribuciones Lambda generalizadas (GLD's) para modelizar las distribuciones de puntuación de correspondencias espectro-péptido. La función Lambda generalizada puede definirse mediante su distribución percentil:

$$Q(y) = Q(y, \lambda_1; \lambda_2; \lambda_3; \lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2}, \qquad (1)$$

donde $0 \leq y \leq 1$. Los parámetros $\lambda 1$y $\lambda 2$ son, respectivamente, los parámetros de localización (entendido como el desplazamiento de la distribución en el eje de abscisas) y de escala (que determina la altura de la distribución), y $\lambda 3$ y $\lambda 4$ determinan, respectivamente, la asimetría de la distribución (respecto a un eje vertical) y su curtosis (definida como el grado de concentración en torno al pico máximo). Una descripción adecuada de las restricciones necesarias en estos parámetros para proporcionar GLD's válidas puede encontrarse, por ejemplo, en la Referencia [7]. A partir de la función percentil, la densidad de probabilidad en x=Q(y) se obtiene como

$$f(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3} + \lambda_4 (1-y)^{\lambda_4 - 1}} \quad (2)$$

Dado que y se define como la probabilidad de que x ≤ Q(y), la modelización de las GLD's a partir de los histogramas de datos observados requiere la conversión de los puntos de datos en una frecuencia de escala relativa, el cálculo del valor de Q(y) para todos los puntos y el agrupamiento de los puntos de datos de acuerdo a dicho valor. Con el objetivo de ajustar las GLD's a los histogramas de datos, se emplea el método de percentiles descrito en la Referencia [7], en el que se calculan cuatro muestras estadísticas empleadas como estimadores de los parámetros de la distribución. De entre todos los conjuntos de parámetros (λ1, λ2, λ3, λ4) compatibles con el conjunto de estimadores obtenidos para cada histograma, se selecciona la GLD que mejor se ajusta a los datos observados como aquélla que minimiza el indicador de error contemplado, definido este último por medio de la expresión

$$\sum_{i=1}^{K} (y_i - f_i)^2 \quad (3)$$

donde y_i es el valor observado en la i-ésima casilla del histograma de puntuaciones (con K casillas) y f_i es el valor que predice el modelo GLD en consideración (densidad de probabilidad), de forma similar a un ajuste por mínimos cuadrados.

3. Estimación de valores-p y de tasas de error en la identificación de péptidos: Como consecuencia de que no existe una expresión cerrada para la función de probabilidad del tipo y = F(x), el conjunto de los valores-*p* asociados a cada punto de los datos se calcula numéricamente. Dado un conjunto de valores-p asociados a los péptidos y clasificados en orden ascendente, la proporción esperada de observaciones de datos que superan un umbral de valor-*p* p_i depende del volumen de los datos, así como del número i de puntos que poseen igual o mayor valor-p. Esta cantidad, denominada como tasa de falsa detección (FDR), da una medida del error esperado:

$$FDR_i = \frac{Np_i}{i} \quad (4).$$

Las tasas de error también pueden ser estimadas mediante búsquedas en bases de datos de secuencias híbridas diana/señuelo, contando el número de impactos señuelo que superan un determinado umbral de valor-p. Este valor, calculado a partir de la proporción de identificaciones señuelo observadas entre todas las identificaciones realizadas para un filtro dado, se denomina tasa de impacto en señuelo (DHR) y se define como

$$DHR_i = \frac{\alpha D_i}{i} \quad (5)$$

donde D_i es el número de asignaciones a péptidos señuelo con un valor-*p* igual o inferior a p_i. El parámetro α varía en función del tipo de base de datos de secuencias empleada. Para bases de datos híbridas diana/señuelo con secuencia invertida, α es igual a 2.

Otras realizaciones de la presente invención pueden incluir el uso de otras medidas de estimación del error como, por ejemplo, la probabilidad proporcionada por la "Familywise error rate" (FWER), definida como

$$FWER_i = 1 - (1 - p_i)^N \quad (6)$$

donde pi es el i-ésimo mejor valor-p, de entre N valores-p obtenidos.

4. Cálculo de los valores-p y puntuaciones de identificación de proteínas y tasas de error: Las asignaciones secuencia-péptido se agrupan dentro de una secuencia de proteína precursora. De los valores-p de un número h dado de iones candidatos, asignados a una proteína dada, el valor de la puntuación de la proteína se define como

$$S_p = \sum_{i=1}^{h} -\log(p_i) \quad (7)$$

donde p_i son los valores-*p* de los iones candidatos calculados en los modelos GLD correspondientes. Opcionalmente, el valor de la puntuación de la proteína también puede definirse como la suma de las meta-puntuaciones de péptido. Del mismo modo, los valores de las FDR y DHR se calculan de la forma descrita en el punto anterior para cada grupo de similitud (definido este término como el conjunto de proteínas que comparten al menos un péptido identificado), tomando como valor-*p el valor-p de proteína* más pequeño dentro del grupo.

5. Integración de datos de múltiples motores de búsqueda y cálculo de meta-puntuaciones: La estrategia de integración de datos de múltiples motores de búsqueda se representa esquemáticamente en la Figura 1. Los espectros MS/MS se asignan a secuencias de péptidos mediante el uso de múltiples motores de búsqueda de secuencias (meta-búsqueda). Ejemplos actuales de estos motores son, por ejemplo, las aplicaciones MASCOT (distribuido por Matrix Science Inc.), X!TANDEM (distribuido por The Global Proteome Machine Organization), OMSSA (distribuido por el National Center for Biotechnology Information) o InsPect (distribuido por el Center for Computacional Mass Spectrometry), entre otros. Tras la identificación de secuencias a los péptidos candidatos, se ajustan las GLD's y se calculan todos los valores-*p* con sus correspondientes puntuaciones, del modo descrito en los párrafos anteriores. En una realización preferente de la presente invención se construye una tabla que contenga la máxima puntuación obtenida por cada motor de búsqueda para cada espectro MS/MS en el conjunto de datos. Con esta información, se define la meta-puntuación de un espectro *j* dado de un conjunto de datos como

$$S_j = \arg\max_k (GLD(1-p_{jk}, 0, 0.2142, 0.1488, 0.1488) + \beta A_{jk}) \qquad (8),$$

donde se toma el valor de k que maximiza el valor de la puntuación S_j para un espectro dado. La variable p_{jk} es el valor-p calculado por medio del modelo GLD correspondiente a un motor de búsqueda k dado, asociado a un péptido candidato. La función de distribución GLD(1-pjk,0,0.2142,0.1488,0.1488) es el valor de la función percentil (definida como la función inversa de la distribución acumulada) de la GLD en el valor-p p_{jk}, de forma que se obtenga aproximadamente una distribución normal, siempre y cuando los valores-p se distribuyan uniformemente. A_{jk}, definido como el parámetro de concordancia del motor de búsqueda, indica el número de otros motores de búsqueda que han proporcionado el mismo péptido candidato que el k-ésimo motor, para el j-ésimo espectro. Por último, β es un coeficiente cuyo valor ha de ser optimizado específicamente en cada conjunto de datos, seleccionando aquel valor que maximice el número de espectros recuperados para un valor dado de la DHR. El valor óptimo del coeficiente de concordancia también puede ser estimado mediante un método numérico distinto, empleando una formulación más compleja para bonificar la concordancia entre motores, en lugar de asumir una dependencia lineal entre el número de concordancia y la magnitud de la bonificación.

En una segunda realización preferente de la invención se lleva cabo un procedimiento por el cual, para un espectro dado j, en lugar de tomar el mejor candidato de cada motor, se toman los I mejores candidatos, ordenados de mayor a menor puntuación (i=1,…,I). Posteriormente se define un parámetro de concordancia extendido A_{ijk}, que designa el número de otros motores (k=1,…,K) que proporcionan como mejor candidato (i=1) el mismo péptido que el i-ésimo candidato del k-ésimo motor. Además, cierto número de parámetros accesorios X_1, …, X_n, que representan la contribución de n fuentes adicionales de información, comprendiendo estas fuentes de información uno o más de lo siguiente:

a) Fuentes de información relacionadas con las características fisiomecánicas de las secuencias de péptido candidatas:

- Error del valor m/Z del ion precursor: error absoluto en la medición de la relación masa/carga del ion precursor del espectro de fragmentación en consideración, en valor absoluto, dada una secuencia de péptidos candidata. El cálculo del valor esperado de la relación masa/carga del ion precursor es trivial a partir de la secuencia de péptidos candidata de la carga estimada del ion precursor.

- Error de tiempo de retención: error absoluto del tiempo de retención del espectro de fragmentación en consideración, en valor absoluto. Se aplica cuando los datos se han obtenido usando cromatografía de fase inversa (RPC) acoplada con espectrometría de masas.

- Error de tiempo de retención de fragmentación: error absoluto del tiempo de retención (en la etapa de pre-fragmentación del péptido) del espectro de fragmentación en consideración, en valor absoluto. Se aplica cuando los datos por fragmentación de péptidos se han obtenido a través de cualquier método bioquímico adecuado (intercambio de iones, inversión de fase en PH alcalino, isoelectroenfoque, etc…) acoplada antes de la cromatografía de fase inversa (RPC) acoplada con espectrometría de masas. El valor observado para cada espectro puede ser el tiempo de retención en el cual se obtuvo cada fracción, si está disponible una medida de este valor, o simplemente el número de fracción (el cual, de hecho, es una transformada del orden del valor previo).

b) Fuentes de información relacionadas con el comportamiento esperado de compuesto químico o enzima que ha generado los péptidos analizados por espectrometría de masas:

- Número de dianas internas: número de sitios de corte de la enzima o agente químico que contiene la secuencia de péptidos candidata. Se define una variable binaria para cada valor del número de dianas internas observadas en el experimento, cuyo valor es 1 si este número concuerda con el número de dianas dentro de la secuencia de péptidos candidata, y en otro caso es cero.

- Número de extremos específicos: número de extremos de la secuencia de péptidos candidata cuya secuencia es compatible con el comportamiento esperado del agente químico o enzima que ha generado los péptidos. Se define una variable binaria para cada valor del número de extremos específicos observados en el experimento, cuyo valor es 1 si este número concuerda con el número de dianas dentro de la secuencia de péptidos candidata, y en otro caso es cero.

c) Fuentes de información relacionadas con la generación de espectros múltiples a partir de un único péptido:

- Formas alternativas-carga eléctrica: número de cargas eléctricas diferentes (provistas por el motor como mejores candidatas para un espectro dado), de las que se ha provisto el péptido candidato en el experimento. La disparidad de cargas eléctricas de un péptido dado depende del mecanismo de ionización.

- Formas alternativas-firmas isotópicas: número de configuraciones de firmas isotópicas estables (provistas por el motor como mejores candidatas para un espectro dado), que se han detectado en la secuencia de péptidos candidata en el experimento, cuando los datos vienen de experimentos de etiquetado de isótopos estables (etiquetado de isótopos estables, -SILE).

- Formas alternativas-modificaciones químicas: número de formas de modificación química (provistas por el motor como mejores candidatas para un espectro dado), que se han detectado en la secuencia de péptidos candidata en el experimento, cuando los péptidos pueden experimentar cambio químico durante el proceso de análisis, cuando tales cambios son inducidos por el usuario o no.

- Formas alternativas-mecanismos de fragmentación: número de mecanismos de transformación de iones que han generado espectro (provistos por el motor como mejores candidatos para un espectro dado), a través de los cuales se ha detectado la secuencia de péptidos candidata en el experimento, cuando el experimento combina datos obtenidos usando diferentes mecanismos de fragmentación (por ejemplo, disociación inducida por colisión (CID) o disociación de transferencia de electrones (ETD)).

d) Fuentes de información relativas a características específicas del motor y el rendimiento dependiendo del tipo de datos:

- Carga eléctrica del ion precursor: se define una variable binaria para cada valor de carga eléctrica observado en el experimento, cuyo valor es 1 si la carga eléctrica del ion precursor es igual a dicha carga eléctrica y, en otro caso ,0. Se usa para promover o penalizar formas de carga eléctrica para las cuales el rendimiento de un motor dado es particularmente bueno o malo.

- Mecanismo de fragmentación: se define una variable binaria para cada mecanismo de fragmentación de iones usado en el experimento, cuyo valor es 1 si el espectro en consideración se ha obtenido por dicho mecanismo de fragmentación y, en otro caso, 0. Se utiliza para promover o penalizar mecanismos de fragmentación para los cuales el rendimiento de un motor dado es particularmente bueno o malo.

- Puntuación Delta y puntuaciones adicionales: se define una puntuación delta genérica para todos los motores como la puntuación dada por un motor a una secuencia de péptidos candidata menos la puntuación más alta observada entre los restantes candidatos para el mismo espectro con una puntuación inferior. Esta puntuación es similar a puntuaciones diferenciales similares usualmente llamadas “delta”, que proporcionan algunos motores como SEQUEST. Se define puntuación adicional como cualquier cantidad que es capaz de ser usada como puntuación, y que es proporcionada por el motor junto con la puntuación principal, aunque usualmente es mucho menos informativa que ésta. Por ejemplo, puntuación PRM media, puntuación PRM total, variables de fracción Y y fracción B, provistas por el motor Inspect pueden ser definidas como tales variables, junto con su puntuación principal, llamada puntuación MQ.

e) Fuentes de información relacionadas con la proteína precursora de los péptidos candidatos:

- Proteína precursora (experimento completo): establece una relación entre el número de péptidos con los cuales se ha identificado la proteína (desde todo el espectro del experimento) de una secuencia de péptidos candidata dada y la longitud de la secuencia de dicha proteína. Para este propósito, las proteínas son primeramente ordenadas de mayor a menor número de péptidos identificados, y en segundo lugar por orden decreciente de longitud de cadena y entonces se usa en ambos casos el ranking relativo para generar, a través de una función inversa normal estandarizada, variables que siguen una distribución normal estándar. La diferencia entre estas dos variables se toma como la puntuación de la proteína.

- Proteína precursora (fracción de experimento): igual que anteriormente, pero se cuentan el número de péptidos de la misma proteína en el espectro de una cierta fracción del experimento y no del experimento completo. Puede usarse cuando se ha realizado un fraccionamiento de proteínas, usando cualquier técnica bioquímica apropiada, antes de la generación de péptidos a analizar por espectrometría de masas.

- Proteína precursora (agrupamiento): se cuenta el número de fracciones diferentes en el experimento, en las que aparecen las secuencias de péptidos candidatas en cuestión. Se toma un gran número K de muestras aleatorias (p.ej. K=1000) de péptidos identificados en el experimento de tamaño N, donde N es el número de péptidos identificados en la proteína precursora y se cuenta el número de fracciones k_s diferentes de los péptidos en la muestra. Se cuenta el número de muestras aleatorias diferentes R en las que k_s adopta un valor mayor que k_t y se define la fuente de información para el agrupamiento de la proteína precursora como R/K. Puede usarse cuando se ha hecho un fraccionamiento de proteínas, usando cualquier técnica bioquímica apropiada, antes de la generación de péptidos a ser analizados por espectrometría de masas.

Es posible, además, el uso de transformaciones numéricas de las fuentes adicionales de información mencionadas anteriormente, sean éstas transformaciones de orden, transformaciones no lineales, categorías arbitrarias basadas en rangos de valores, probabilidades o densidades de probabilidad calculadas a partir de estas fuentes adicionales de información, reemplazando éstas o en combinación con éstas, usando estas transformaciones también como fuentes adicionales de información.

Después de determinar qué fuentes de información se usan, se define, de este modo, la puntuación extendida del i-ésimo candidato proporcionado para el j-ésimo espectro por el k-ésimo motor como:

$$s_{ijk} = GLD(1-p_{ijk}, 0, 0.2142, 0.1488, 0.1488) + \beta_1 x_{1ijk} + \ldots + \beta_n x_{nijk} + \gamma A_{ijk}) \quad (9).$$

donde, p_{ijk} se calcula como se ha descrito anteriormente para todos los candidatos de cada motor a partir de la puntuación proporcionada por dicho motor, los coeficientes $\beta_1,\ldots\beta_n$ y γ se optimizan mediante cualquier método matemático de optimización en varias dimensiones, por ejemplo maximizando el número de espectros o péptidos recuperados fijando un determinado umbral de DHR. En cada iteración del método de optimización se reordenan de mayor a menor valor de s_{ijk} los I mejores candidatos de los K motores y se les reasigna el índice i con el objetivo de recalcular los valores A_{ijk}. Finalmente, se define la meta-puntuación del j-ésimo espectro como:

$$S_j = \arg\max_{i,k}(s_{ijk}) \quad (10)$$

donde i,j, y k son número enteros, tomando como péptido candidato para el j-ésimo espectro el i-ésimo candidato del k-ésimo motor, tal que los valores de i y k maximicen el valor de S_j.

En una tercera realización preferente de la invención, se define el parámetro de concordancia en su forma ponderada de la siguiente manera:

$$A_{ijk} = \sum_{l=1, l\neq k}^{k} w_{kl} a_{ijkl}, \quad (11)$$

donde a es la matriz de variables binarias de tamaño KxK que indica cuáles de los K motores proporcionan el mismo péptido candidato que el k-ésimo motor, y w una matriz con coeficientes de peso de las concordancias entre motores. Nótese que fijando a 1 todos los valores de la matriz w se obtiene la meta-puntuación de la ecuación 9, y fijando i=1 además, se obtiene la meta-puntuación de la ecuación 8. El valor de estos coeficientes podría calcularse, por ejemplo, a partir de las frecuencias de concordancia entre motores observadas en los péptidos señuelo, o bien asumiendo un mismo valor inicial para todos ellos (p.ej., 1/(K(K-1), y optimizando a continuación dichos valores según lo descrito para la ecuación 9.

En una cuarta realización preferente de la invención, se asigna un coeficiente de peso distinto a cada motor durante la fase de meta-puntuación, siendo éste establecido o calculado a priori, de modo que se pueda incorporar la posibilidad de favorecer los resultados obtenidos por unos motores sobre otros, si las particularidades del experimento analizado lo requiriesen. El valor de estos coeficientes podría calcularse de modo análogo a los descritos anteriormente.

En una quinta realización preferente de la invención, después de generar meta-puntuaciones para cada motor, se establece un orden de integración de fuentes adicionales de información, de modo que, para un motor dado, se

incorpora a la meta-puntuación una fuente de información individual adicional, ignorando la información de concordancia con otros motores en la ecuación 9, y se optimiza su coeficiente β usando un método numérico de optimización en solo una dimensión. Después de obtener una nueva meta-puntuación a través de este proceso, se toma una nueva fuente de información. Este proceso se repite hasta que todas las fuentes adicionales de información se han incorporado a la meta-puntuación. La ventaja de este proceso de meta-puntuación que incorpora fuentes adicionales de información en etapas es que tiene las propiedades teóricas necesarias para eliminar cualquier correlación entre fuentes adicionales de información. Después de actualizar la meta-puntuación de todos los motores, se incorpora la información de concordancia, usando los métodos descritos en las ecuaciones 8 o 9, y después el método descrito en la Ecuación 10.

En una sexta realización preferente de la invención, se establece un orden de integración de los distintos motores de búsqueda, de modo que el proceso empieza con dos motores (preferiblemente aquellos dos que proporcionan la mayor sensibilidad, por ejemplo definidos como el número de identificaciones y una tasa de error dada) y se aplican las ecuaciones 8 o 9 y la reivindicación 10. El resultado de este nuevo proceso es como un nuevo motor de “consenso”, entonces este resultado se toma junto con el tercer motor y se aplican las ecuaciones 8 o 9 y la reivindicación 10. El proceso se repite hasta que todos los motores se han incorporado al “consenso” preferiblemente en orden de sensibilidad descendiente. La ventaja de este proceso de meta-puntuación por etapas es que tiene las propiedades teóricas necesarias para eliminar cualquier correlación entre motores.

Después de que las secuencias de los péptidos candidatos han sido asignadas a todos los espectros MS/MS, se elimina la redundancia, manteniendo, para cada combinación de secuencia de péptido, carga eléctrica y patrón de estructura química, aquélla que posee la mayor meta-puntuación. Posteriormente, se obtiene una distribución única de meta-puntuación para cada conjunto de datos, ya que las meta-puntuaciones son independientes del estado de carga del ión precursor. A partir de los valores-p obtenidos del modo descrito en los puntos anteriores, se pueden calcular tanto las tasas de error FDR y DHR (en los dos niveles, péptido y proteína), como los valores-p para proteínas.

Entre las ventajas del método de meta-búsqueda descrito por la presente invención respecto a otros métodos de búsqueda de secuencias conocidos, cabe señalar los siguientes:

- Es un método completamente generalizable para su aplicación a cualquier número de motores de búsqueda.

- Emplea un método estándar para obtener las funciones de distribución estadística, aplicable a los resultados de cualquier motor de búsqueda.

- Emplea una modelización estadística robusta que permite la elección de una única combinación secuencia de péptidos, estado de carga y patrón de estructura química por espectro.

- El método de meta-búsqueda y su sistema de meta-puntuación agrega información adicional que no puede ser obtenida mediante la búsqueda con un solo motor.

- Integra en su formulación el empleo de parámetros de concordancia, definidos como el número de otros motores de búsqueda que han proporcionado el mismo péptido candidato que un motor dado.

- En cuanto a la detección de proteínas, se emplea un método estadístico riguroso, no sesgado, que emplea un filtrado FDR.

- Adicionalmente, el método reivindicado permite incorporar otras fuentes de información adicionales a la concordancia del motor, tales como el error de masa del péptido precursor, el error en el tiempo de retención, la especificidad de la digestión enzimática o la concordancia con la secuenciación de novo de la información. Esta flexibilidad permite al método de meta-búsqueda la integración de datos empleando diferentes preparaciones de muestras, métodos de digestión de proteínas y mecanismos de fragmentación de iones.

- A modo de ejemplo, se incluyen aquí los resultados de los ensayos realizados mediante el método reivindicado por la presente invención (Véanse las tablas 1a-1f y tablas 2a-2d), para las muestras de datos de acceso público RaftFlow, PAe000038-39 (disponible en la página web PeptideAtlas), PAe000114 (también en PeptideAtlas), iPRG2008 (del Association of Biomolecular Resource Facilities Proteome Informatics Research Group), evaluado para dos conjuntos de parámetros de búsqueda distintos (y distinguidos por los nombres iPRG2008 e iPRG2008-NE). La descripción detallada de estos conjuntos de datos y de sus experimentos asociados puede consultarse en la Referencia [6]. Adicionalmente, se incluyen los resultados a nivel péptido del experimento SKHep-LA-I (Laboratorio de Proteómica, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas), consistente en el enriquecimiento de péptidos que son ligandos naturales de las moléculas del complejo mayor de histocompatibilidad de tipo I (MHC-I). En dicho experimento se purifican los péptidos a partir de células de la línea Sk-Hep, que expresan los alelos de clase I HLA-A*0201, HLA-A*2402, HLA-B*3502 y HLA-B*4403. Estos péptidos son generados por un proceso de digestión natural en el interior de la célula, unidos a moléculas de MHC y transportados a la superficie celular, donde son presentados a las células del sistema inmunitario. Los distintos alelos de los genes que codifican las proteínas MHC pueden tener un repertorio de péptidos ligandos distintos, con propiedades estructurales ligeramente diferentes. Se cree que algunos de estos alelos

están asociados a enfermedades autoinmunes, por lo que disponer de herramientas automatizadas para la caracterización a gran escala de repertorios de moléculas MHC (ya sean de tipo I o II) es de notable interés biomédico. El experimento CID-HLA-ETD es una réplica del experimento anterior, pero los espectros MS/MS se obtuvieron usando dos mecanismos de fragmentación diferentes conocidos por sus iniciales en inglés como CID (disociación inducida por colisión) y ETD (disociación por transferencia de electrones). El experimento de fosfo-péptidos ABRF201 0 corresponde al análisis, también generándose tanto los espectros CID como ETD de los fosfo-péptidos enriquecidos por cromatografía IMAC (cromatografía de afinidad por metales inmovilizados), de una muestra de proteína humana proporcionada por la Asociación de Centros de Investigación Molecular (ABRF). La fosforilación es una modificación traslacional de gran importancia en el proceso de señalización intracelular, de modo que los resultados en la identificación de los fosfo-péptidos por espectrometría de masas aquí mostrados son de gran importancia en el campo de la investigación biomédica y tanto en biotecnología básica como aplicada. El experimento Ecoli SILE-SILAC corresponde al análisis de una muestra de dos poblaciones de la bacteria Escherichia Coli en un cultivo, marcadas con diferentes formas isotópicas del aminoácido lisina (forma nativa o pesada 13Cx6, 15Nx2, +8Da) usando la técnica SILAC (Marcado Isotópico Estable por Aminoácidos en Cultivo Celular), cuyos extractos de proteínas fueron fraccionados por electroforesis en gel de poliacrilato antes de la digestión con tripsina. El experimento “Suero Frac. RP-PH alcalino” corresponde al análisis de una muestra de suero humano en el cual, después de la digestión con tripsina del extracto de proteína, los péptidos obtenidos fueron fraccionados por cromatografía de fase inversa a PH alcalino (aproximadamente 10.9).

El tratamiento de los datos de los diferentes experimentos se ha realizado mediante el uso conjunto de los cuatro motores de búsqueda InsPect, MASCOT, X!TANDEM (utilizado este último en dos versiones de puntuación, clásica y “k-score”) y OMSSA. Los resultados obtenidos por medio del método de meta-búsqueda reivindicado por la presente invención se resumen en la Tabla 1. Para la mayoría de los conjuntos de datos empleados, el sistema de meta-puntuación combinada de todos los motores de cálculo, empleando los valores-p obtenidos mediante modelización GLD, proporciona un incremento sustancial del número de péptidos identificados, comparado con el resultado obtenido individualmente en cualquiera de los motores considerados. Para el caso del experimento PAe000114, dado que está claramente dominado por el resultado del motor InsPect, se incluyen también, a modo de comparación, los resultados de la meta-búsqueda excluyendo dicho motor. La combinación del resto de los motores, incluyendo la información de concordancia, proporcionó una eficiencia 19% superior a la obtenida por OMSSA individualmente, y una eficiencia aún mayor en el resto de motores. En general, el empleo de la información de concordancia mejora la sensibilidad de todos los experimentos, incrementando entre un 9% y un 26% el número de péptidos correctamente identificados (con una FDR ≤ 0.05 sobre un conjunto no redundante). Respecto a la detección de proteínas, el número de identificaciones con dicho umbral de error, aumenta entre un 6% y un 60% después de su clasificación mediante meta-puntuación. Las tablas 2(a-d) muestran los resultados de los procesos de meta-puntuación incorporando por etapas una, ninguna o varias fuentes de información adicional. Como puede verse, todas las fuentes adicionales de información descritas ayudan a incrementar la eficiencia del proceso de meta-puntuación, a juzgar por el significativo aumento de localizaciones espectro-secuencia recuperadas para un cierto valor de la tasa de error, especialmente cuando se usan varias de estas fuentes adicionales de información en combinación. Se hace notar que algunas de estas fuentes adicionales de información están basadas en peculiaridades relacionadas con el diseño experimental que ningún motor de búsqueda es capaz de incorporar a su sistema de puntuación, como desviaciones de los valores esperados del tiempo de retención tiempo de retención durante la fragmentación de péptidos anterior al análisis por espectrometría de masas (experimento Suero Frac. PH alcalino), anterior a la pre-fragmentación de proteínas (experimentos SILE-SILAC), formas alternativas de carga, firmas isotópicas (experimentos SILE-SILAC) o mecanismos de fragmentación (experimentos CID HLA-ETD y fosfo-péptidos ABRF2010), etc. Además, el método descrito permite usar de forma óptima y continuar extrayendo información de esas fuentes incluso en casos en los que el motor ya usa estas fuentes en su sistema de puntuación, como en MASCOT, que internamente usa el error en el valor m/Z del ion precursor para calcular sus puntuaciones (Véanse los datos del experimento Suero Frac. PH alcalino) así como incorporar fácilmente puntuaciones delta y puntuaciones complementarias proporcionadas por el motor además de la puntuación principal (Véanse el experimento fosfo-péptidos ABRF201 0, los datos del motor Inspect y el experimento Suero Frac. A PH alcalino y los datos del motor MASCOT). Bajo estas condiciones, la efectividad del proceso es incluso mayor cuando se usa información de múltiples motores en vez de un solo motor, como se ve claramente en los experimentos de fosfo-péptidos HLA-ETD y CID ABRF 201 0.

TABLAS DE RESULTADOS:

Tabla 1 (más abajo): Comparación entre los resultados de modelado usando un único motor para diferentes experimentos y los resultados usando el método de meta-búsqueda. Los índices usados son: InsPect, K, X-Tandem con “puntuación k”; M, MASCOT, OR, OMSSA, T, X! TANDEM clásico. Las listas separadas por comas corresponden al uso de múltiples motores. “Concord” indica si la información de concordancia se ha tomado en cuenta. “No. Pept” indica el número de concordancias de péptidos no redundantes obtenidos para el filtro FDR (o DHR, si existe) dado. “No. Prot” indica el número de grupos de agregación de proteínas obtenidos para el filtro FDR (o DHR, si existe) dado. “N/a” indica “no aplicable.

Tabla 1a

Experimento F							
Motor(es)	Concord.	Nº Pépt. (FDR ≤ 0.05)	DHR	Nº Pépt. (DHR ≤ 0.05)	Nº Prot. (FDR ≤ 0.05)	Prot. DHR	Nº Prot. (DHR ≤ 0.05)
I	n/a	1751	0,037	1851	410	0,062	360
K	n/a	1897	0,038	1986	455	0,048	456
M	n/a	1565	0,06	1511	411	0,044	422
O	n/a	1720	0,029	1825	447	0,072	430
T	n/a	1545	0,044	1579	426	0,069	412
I,K,M,O,T	no	2489	0,042	2552	527	0,059	515
I,K,M,O,T	sí	2708	0,049	2714	567	0,062	555

Tabla 1b

Experimento PAe00							
Motor(es)	Concord.	Nº Pépt. (FDR ≤ 0.05)	DHR	Nº Pépt. (DHR ≤ 0.05)	Nº Prot. (FDR ≤ 0.05)	Prot. DHR	Nº Prot. (DHR ≤ 0.05)
I	n/a	522	0,042	572	150	0,053	108
K	n/a	455	0,105	58	91	0,042	96
M	n/a	521	0,054	505	130	0,046	145
O	n/a	616	0,058	579	182	0,062	154
T	n/a	409	0,058	394	149	0,067	147
I,K,M,O,T	no	807	0,138	443	151	0,066	147
I,K,M,O,T	sí	993	0,13	801	239	0,05	239

Tabla 1c

Experimento PAe							
Motor(es)	Concord.	Nº Pépt. (FDR ≤ 0.05)	DHR	Nº Pépt. (DHR ≤ 0.05)	Nº Prot. (FDR ≤ 0.05)	Prot. DHR	Nº Prot. (DHR ≤ 0.05)
I	n/a	5997	0,031	6503	1274	0,072	1193
K	n/a	2829	0,066	2586	915	0,066	858
M	n/a	4277	0,043	4381	1251	0,059	1211
O	n/a	4713	0,048	4752	1122	0,072	1015
T	n/a	3217	0,028	3513	1179	0,067	1125
I,K,M,O,T	no	5987	0,0281	6674	1356	0,05	1347
I,K,M,O,T	sí	6711	0,0355	7261	1326	0,04	1426
K,M,O,T	no	4765	0,0293	5223	1258	0,06	1197
K,M,O,T	sí	5685	0,0384	5973	1334	0,067	1290

Tabla 1d

Experimento iPRG							
Motor(es)	Concord.	Nº Pépt. (FDR ≤ 0.05)	DHR	Nº Pépt. (DHR ≤ 0.05)	Nº Prot. (FDR ≤ 0.05)	Prot. DHR	Nº Prot. (DHR ≤ 0.05)
I	n/a	148	0,056	145	43	0,045	50
K	n/a	673	0,029	708	255	0,047	258
M	n/a	555	0,053	547	191	0,032	197
O	n/a	497	0,023	576	182	0,046	182
T	n/a	408	0,054	402	164	0,048	168
I,K,M,O,T	no	725	0,0438	727	228	0,064	221
I,K,M,O,T	sí	878	0,0367	913	316	0,059	308
K,M,T	no	712	0,0501	708	235	0,044	235
K,M,T	sí	892	0,0647	877	298	0,069	289

Tabla 1e

Experimento iPRG2							
Motor(es)	Concord.	Nº Pépt. (FDR ≤ 0.05)	DHR	Nº Pépt. (DHR ≤ 0.05)	Nº Prot. (FDR ≤ 0.05)	Prot. DHR	Nº Prot. (DHR ≤ 0.05)
K	n/a	241	0,027	216	154	0,052	152
M	n/a	357	0,074	387	106	0,019	122
T	n/a	85	0,044	91	49	0,041	50
K,M,T	no	341	0,0629	335	149	0,04	153
K,M,T	sí	805	0,0961	708	239	0,025	249

Tabla 1f

Experimento SKHe							
Motor(es)	Concord.	Nº Pépt. (FDR ≤ 0.05)	DHR	Nº Pépt. (DHR ≤ 0.05)	Nº Prot. (FDR ≤ 0.05)	Prot. DHR	Nº Prot. (DHR ≤ 0.05)
K	n/a	36	0,049	41	n/a	n/a	n/a
M	n/a	25	0,2	3	n/a	n/a	n/a
O	n/a	6	0,2	5	n/a	n/a	n/a
T	n/a	14	0,12	9	n/a	n/a	n/a
K,M,O,T	no	56	0,14	3	n/a	n/a	n/a
K,M,O,T	sí	180	0,054	180	n/a	n/a	n/a

Tabla 2 (más abajo). Comparación entre resultados de modelización mediante el uso de un único motor o una combinación de vario motores por etapas (usando información de concordancia), basada en los datos generados por un único mecanismo de fragmentación o a partir de múltiples mecanismos incorporando por etapas una, ninguna o varias fuentes adicionales de información al proceso de meta-puntuación. Los índices empleados son los mismos que en la Tabla 1, a los que se añade P (PHENIX). El rendimiento del proceso se reseña como el número de localizaciones de espectro-secuencia recuperadas al superar una tasa de error particular medidas como DHR (0,01; 0,05 y 0,1). Para fuentes adicionales de información, "TODAS" indica que se incorporaron todas las fuentes de información descritas que estaban disponibles para esos datos; "NINGUNA" indica que no se usó ninguna fuente adicional de información.

. Tabla 2a. Experimento Suero Frac. RP-PH alcalino					
Motor(es)	Mec. Frag	Fuentes adicionales de información	Nº espectros recuperados DHR<=0,01	DHR<=0,05	DHR<=0,1
M	CID	NINGUNA	393	524	619
M	CID	Error de precursor m/Z	444	569	671
M	CID	Error de tiempo de retención	418	548	657
M	CID	Error de fraccionamiento de tiempo de retención	453	576	656
M	CID	Formas alt. Carga eléctrica	424	638	744
M	CID	Dianas internas	420	552	647
M	CID	Puntuación delta	401	525	622
M	CID	Experimento de precursor de proteína	660	798	915
M	CID	TODAS	832	948	1028

Tabla 2b. Experimento Ecoli SILE-SILAC					
Motor(es)	Mec. Frag	Fuentes adicionales de información	Nº espectros recuperados DHR<=0,01	DHR<=0,05	DHR<=0,1
M	CID	NINGUNA	5687	7424	8302
M	CID	Formas alt. Formas isotópicas	6458	7724	8618
M	CID	Experimento precursor de proteína	6458	8122	9125
M	CID	Fracción precursor de proteína	6084	7539	8661
M	CID	TODAS	7750	8911	9637

Tabla 2c. Experimento HLA CID-ETD					
Motor(es)	Mec. Frag	Fuentes adicionales de información	Nº espectros recuperados DHR<=0,01	DHR<=0,05	DHR<=0,1
M	CID,ETD	NINGUNA	4	4	4
M	CID,ETD	Formas alt -Mecanismos fragmentación	169	195	229
K	CID,ETD	NINGUNA	1	1	1
K	CID,ETD	Formas alt -Mecanismos fragmentación	128	215	232
T	CID,ETD	NINGUNA	3	3	3
T	CID,ETD	Formas alt -Mecanismos fragmentación	43	64	76
M,K	CID,ETD	Formas alt -Mecanismos fragmentación	211	322	397
M,K	CID,ETD	Formas alt -Mecanismos fragmentación	256	336	421

Tabla 2d. Experimento Fosfo-péptidos ABRF2010					
Motor(es)	Mec. Frag	Fuentes adicionales de información	Nº espectros recuperados DHR<=0,01	DHR<=0,05	DHR<=0,1
I	CID	NINGUNA	3	3	3
I	CID	Dianas internas	4	4	4
I	CID	Formas alt –carga eléctrica	6	6	6
I	CID	Error precursor m/Z	9	9	9
I	CID	Extremos específicos	11	11	11
I	CID	Puntuaciones complementarias	18	18	18
I	CID	Experimento precursor proteína	15	15	15
I	CID	TODAS	20	20	21
K	CID,ETD	NINGUNA	8	8	8
K	CID,ETD	Formas alt -Mecanismos fragmentación	26	26	30
O	CID,ETD	NINGUNA	2	2	2
O	CID,ETD	Formas alt -Mecanismos fragmentación	21	21	25
M	CID,ETD	NINGUNA	13	13	13
M	CID,ETD	Formas alt -Mecanismos fragmentación	26	26	28
T	CID,ETD	NINGUNA	10	10	10
T	CID,ETD	Formas alt -Mecanismos fragmentación	15	15	15
P	CID,ETD	NINGUNA	13	13	13
P	CID,ETD	Formas alt -Mecanismos fragmentación	26	26	29
M, P	CID,ETD	Formas alt -Mecanismos fragmentación	35	35	36
M, K	CID,ETD	Formas alt -Mecanismos fragmentación	31	31	34
M, O	CID,ETD	Formas alt -Mecanismos fragmentación	27	27	36
M, T	CID,ETD	Formas alt -Mecanismos fragmentación	22	23	23
K, T	CID,ETD	Formas alt -Mecanismos fragmentación	25	25	31
O, T	CID,ETD	Formas alt -Mecanismos fragmentación	20	20	22
O, P	CID,ETD	Formas alt -Mecanismos fragmentación	26	26	34
K, P	CID,ETD	Formas alt -Mecanismos fragmentación	32	32	42
K, O	CID,ETD	Formas alt -Mecanismos fragmentación	29	29	32
P, T	CID,ETD	Formas alt -Mecanismos fragmentación	27	27	30
M, I	CID,ETD	Formas alt -Mecanismos fragmentación	20	20	25
K, I	CID,ETD	Formas alt -Mecanismos fragmentación	18	18	20
O, I	CID,ETD	Formas alt -Mecanismos fragmentación	24	24	33
P, I	CID,ETD	Formas alt -Mecanismos fragmentación	16	16	22
T, I	CID,ETD	Formas alt -Mecanismos fragmentación	16	16	16
O, T, I	CID,ETD	Formas alt -Mecanismos fragmentación	20	24	24
M, P, O	CID,ETD	Formas alt -Mecanismos fragmentación	35	35	36
M, P, I	CID,ETD	Formas alt -Mecanismos fragmentación	21	21	25

REIVINDICACIONES

1. Método de identificación de péptidos y proteínas a partir de datos de espectrometría de masas y búsqueda en bases de datos de secuencias empleando, al menos, dos motores diferentes de búsqueda, en el que se obtienen modelos de distribución de puntuaciones de identificación espectro-péptido para péptidos candidatos identificados por cada uno de dichos motores y se asigna un valor de probabilidad o una tasa de error a partir de estos modelos a cada puntuación, en el que:

a) Se modelizan las puntuaciones de identificación espectro-péptido para péptidos candidatos calculadas en cada motor por medio de funciones de distribución Lambda generalizadas (GLD), calculando los valores de probabilidad de las identificaciones espectro-péptido;

y caracterizado porque

b) Se calcula la contribución del solapamiento a los resultados del péptido candidato entre los distintos motores utilizados, usando parámetros de concordancia de las identificaciones péptido-secuencia, en el que esos parámetros coincidentes se definen como el número de motores de búsqueda que proporcionan el mismo candidato péptido provisto por otros motores;

c) para el grupo de todos los motores de búsqueda usados, se construyen meta-puntuaciones de identificación espectro-péptido a partir de los valores de la probabilidad de identificación espectro-péptido de la etapa (a) y los parámetros de concordancia usados en la etapa (b);

d) Las meta-puntuaciones construidas en la etapa (c) se modelizan utilizando funciones de distribución Lambda generalizadas (GLD), obteniendo la el valor de probabilidad de la identificación espectro-péptido o la tasa de error para obtener una clasificación estadística de la identificación espectro-péptido, para el grupo de todos los motores de búsqueda usados.

2. Método según la reivindicación 1, en el que las bases de datos de secuencias utilizadas son bases de datos híbridas diana o señuelo.

3. Método según cualquiera de las reivindicaciones 1-2, caracterizado porque la tasa de error empleada viene dada por la tasa de falsa detección (FDR).

4. Método según cualquiera de las reivindicaciones 1-2, caracterizado porque la tasa de error empleada viene dada por la tasa de impacto en señuelo (DHR), o por la probabilidad de obtener al menos un falso positivo (FWER).

5. Método según cualquiera de las reivindicaciones 1-4, caracterizado porque se asigna un coeficiente de peso distinto a cada motor durante la fase de meta-puntuación, siendo éste establecido *a priori* o calculado en función de cualquier característica de los motores y/o las bases de datos de secuencias empleados, por la cual se puede favorecer los resultados de unos motores frente a otros.

6. Método según cualquiera de las reivindicaciones 1-5, caracterizado porque se establece una relación entre la meta-puntuación calculada para una identificación espectro-péptido y las características de la secuencia del péptido candidato, tales como su longitud, presencia o ausencia de sub-secuencias o motivos estructurales, o concordancia de la secuencia del péptido con lo esperado a partir del mecanismo de corte del agente químico utilizado en la digestión de las proteínas.

7. Método según cualquiera de las reivindicaciones 1-6, caracterizado porque se establece una relación entre la meta-puntuación calculada para una identificación espectro-péptido y otras variables medibles, como el error observado en la medición de la masa del precursor, la movilidad iónica, la predicción del tiempo de retención durante la separación cromatográfica, la predicción del punto isoeléctrico en una posible separación por isoelectroenfoque, o medidas similares obtenidas a partir de variantes de estas técnicas, o transformaciones de dichas medidas.

8. Método según la reivindicación 7, caracterizado porque se obtiene, para cada espectro, las mejores puntuaciones de cada uno de los motores, se somete a dichas puntuaciones a meta-puntuación, utilizando una o más de las siguientes fuentes de información adicionales:

- relativas a las características físico-químicas de las secuencias de péptidos candidatas tales como el error m/z en el ion precursor, el error en el tiempo de retención o el error en el tiempo de retención de fragmentación;

- relativas al comportamiento esperado del agente químico o la enzima que ha generado los péptidos analizados por espectrometría de masas, tales como el número de dianas internas o el número de extremos específicos;

- relativas a la generación de múltiples espectros a partir del mismo péptido, tales como formas alternativas de carga eléctrica, las firmas isotópicas, modificaciones químicas o mecanismos de fragmentación;

- relativas a las características del motor específico y a su rendimiento dependiendo del tipo de datos, tales como la carga eléctrica del ion precursor, el mecanismo de fragmentación, puntuaciones delta o puntuaciones adicionales;

- relativas al precursor de proteína o a los candidatos péptidos, tales como el precursor de proteína en un experimento completo, el precursor de proteína en una fracción de un experimento o el precursor de proteína por agrupamiento;

y se reordenan las meta-puntuaciones de cada motor, tomando la mejor meta-puntuación de cada motor para completar la fase de meta-puntuación.

9. Método según las reivindicación 8, caracterizado porque se usan las transformaciones numéricas de las fuentes adicionales de información, ya sean transformadas de orden, o transformadas no lineales de densidades de probabilidad calculadas a partir de estas fuentes adicionales de información, bien sustituyéndolas o bien en combinación con ellas, usando estas transformaciones como fuentes adicionales de información.

10. Método según cualquiera de las reivindicaciones 1-9, caracterizado porque se hace uso de parámetros de concordancia extendidos, definidos como el número de otros motores que proporcionan, como mejor candidato, el mismo péptido que uno de los candidatos dados proporcionado por un motor.

11. Método según cualquiera de las reivindicaciones 1-10, caracterizado porque se define el parámetro de concordancia en su forma ponderada mediante el uso de coeficientes de peso.

12. Método según cualquiera de las reivindicaciones 1-11, caracterizado porque distintas búsquedas efectuadas sobre la misma colección de espectros utilizando distintas combinaciones de parámetros configurables del mismo motor se tratan como búsquedas efectuadas por motores distintos.

13. Método según cualquiera de las reivindicaciones 1-12, en el que se establece una relación entre la meta-puntuación calculada para una identificación espectro-péptido e información estructural obtenida mediante interpretación *de novo* del espectro MS/MS.

14. Método según cualquiera de las reivindicaciones 8 -9, caracterizado porque las fuentes adicionales de información se integran en etapas en el proceso de meta-puntuación, generando meta-puntuaciones para cada uno de los motores y estableciendo un orden de integración de dichas fuentes adicionales de información, de forma que para un motor dado se incorpora una fuente individual adicional de información de la meta-puntuación, ignorando la información de concordancia con otros motores, y se obtiene una nueva meta-puntuación, repitiendo este proceso hasta que todas las fuentes adicionales de información se han incorporado a la meta-puntuación y posteriormente añadiendo la información de concordancia.

15. Método según cualquiera de las reivindicaciones 1 -14, en el que los distintos motores son integrados en etapas en el proceso de meta-puntuación, estableciendo un orden de integración de los distintos motores de búsqueda, empezando con la integración de dos motores, y tratando la meta-puntuación en este proceso como un nuevo motor “de consenso”, tomando posteriormente este resultado e integrándolo con un tercer motor, repitiendo sucesivamente el proceso hasta que todos los motores usados en el proceso han sido incorporados al “consenso”.

16. Dispositivo destinado al análisis de resultados de espectrometría de masas en tándem, comprendiendo dicho dispositivo medios orientados a conformar un sistema de análisis de datos para la identificación de péptidos y proteínas, caracterizado porque implementa un método según cualquiera de las reivindicaciones 1-15.

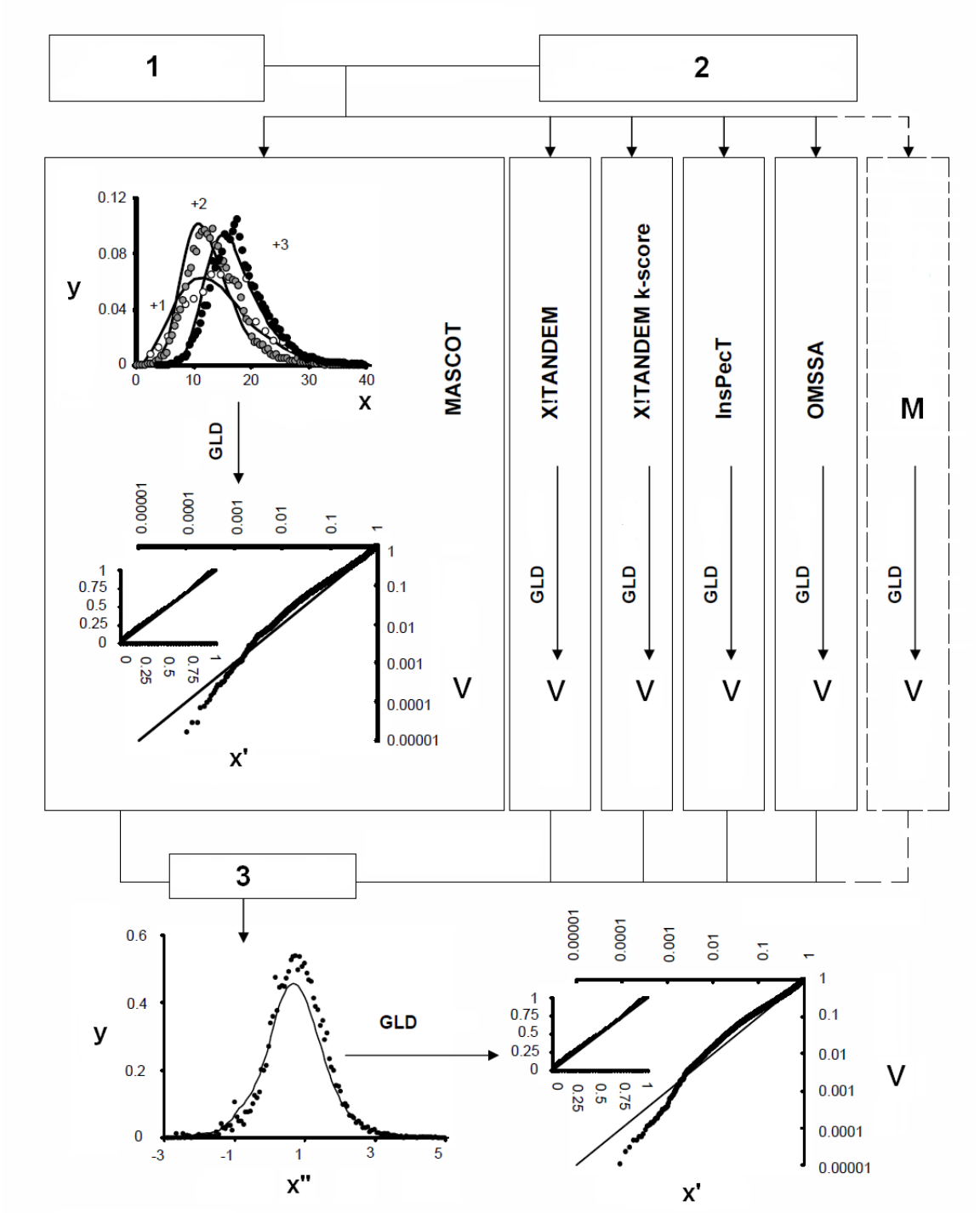
1
2
3
MASCOT
X!TANDEM
X!TANDEM k-score
InsPecT
OMSSA
M
GLD
V
y
x
x'
x''
+1
+2
+3

FIG. 1

REFERENCIAS CITADAS EN LA DESCRIPCIÓN

La lista de referencias citadas por el solicitante es, únicamente, para conveniencia del lector. No forma parte del documento de patente europea. Si bien se ha tenido gran cuidado al compilar las referencias, no pueden excluirse errores u omisiones y la OEP declina toda responsabilidad a este respecto.

Literatura no patente citada en la descripción

- **ROHRBOUGH, JG ; BRESCIA, L. ; MERCHANT, N. ; MILLER, S. ; HAYNES, PA.** Verification of single-peptide protein identifications by the application of complementary database search programs. *J. Biomol. Tech,* 2006, vol. 5, 327-332 **[0009]**
- **HIGGS, RE ; KNIERMAN, MD ; FREEMAN AB ; GELBERT, LM ; PATIL, ST ; HALE, JE.** Estimating the Statistical Significance of Peptide Identifications from shotgun proteomics experiments. *J. Proteome Res,* 2007, vol. 6, 1758-1767 **[0009]**
- **BC SEARLE ; M. TURNER ; A. NESVIZHSKII.** Improving sensitivity by Combining results from multiple MS / MS Search Methodologies. *J. Proteome Res,* 2008, vol. 7, 245-253 **[0009]**
- **ALVES, G. ; WU, WW ; WANG, G. ; SHEN, R.-F. ; YU, Y.-K.** Enhancing confidence peptide identification by combining search methods. *J. Proteome Res,* 2008, vol. 8, 3102-3113 **[0009]**
- **YU W. ; TAYLOR JA ; DAVIS MT ; BONILLA LE ; LEE KA ; AUGER PL ; FARNSWORTH CC ; WELCHER AA ; PATTERNSON SD.** Maximizing the sensivity and reliability of peptide identification in large-scale proteomic experiments by Harnessing multiple search engines. *Proteomics,* 2010, vol. 10, 1172-1189 **[0009]**
- **RAMOS-FERNANDEZ, A. ; PARADELA, A. ; NAVAJAS, R. ; ALBAR, JP.** Generalized method for probability-based peptide and protein identification from tandem mass spectrometry data and sequence database searching. *Mol. Cell. Proteomics,* 2008, vol. 7, 1748-1754 **[0009]**
- **KARIAN ZA ; DUDEWICZ, EJ.** Fitting statistical distributions: the Generalized Lambda Distribution and Generalized Bootstrap Methods. Chapman and Hall / CRC, 2000 **[0009]**