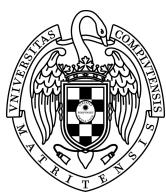

Desarrollo de herramientas bioinformáticas aplicadas a proteómica de alto rendimiento y proteómica dirigida



TESIS DOCTORAL

Vital Vialás Fernández

**Departamento de Microbiología II
Facultad de Farmacia
Universidad Complutense de Madrid**

2015

Desarrollo de herramientas bioinformáticas aplicadas a proteómica de alto rendimiento y proteómica dirigida

Memoria que presenta para optar al título de Doctor
Vital Vialás Fernández

Dirigida por la Doctora
Concha Gil García

Departamento de Microbiología II
Facultad de Farmacia
Universidad Complutense de Madrid

2015

Dedicatoria

*Science is not only compatible with spirituality,
it is a profound source of spirituality.*

Carl Sagan

Agradecimientos

*Frase célebre que vaya bien en los
agradecimientos*

Gracias y bla bla

Resumen

...

...

Índice

Agradecimientos	IX
Resumen	XI
I INTRODUCCIÓN	1
1. Introducción	3
1.1. Proteómica. Conceptos generales	4
1.2. Espectrometría de masas	5
1.2.1. Consideraciones sobre unidades empleadas en espectrometría de masas	6
1.2.2. Componentes de un espectrómetro de masas	7
1.2.3. Espectrometría de masas en Tandem. MS/MS	13
1.3. Digestión de proteínas en péptidos	16
1.4. Proteómica en gel	17
1.4.1. Huella Peptídica	18
1.5. Proteómica a gran escala	19
1.5.1. Separación de péptidos y proteínas sin gel	19
1.6. Asignación Péptido-Espectro	22
1.6.1. Búsqueda en bases de datos de secuencias	22
1.6.2. Otras estrategias de asignación péptido-espectro	28
1.7. Evaluación estadística de las asignaciones PSM	29

1.7.1. Conceptos estadísticos básicos	29
1.7.2. Puntuaciones basadas en distribuciones de espectro individual y promedio	31
1.7.3. Bases de datos señuelo y Tasa de Falsos Descubrimientos (FDR)	34
1.7.4. Modelos mixtos de probabilidad. Probabilidad Posterior	37
1.8. Inferencia de proteínas a partir de péptidos	41
1.9. Herramientas adicionales de post-procesamiento y validación a nivel de péptido y proteína	43
1.10. Proteómica dirigida. SRM/MRM	45
1.11. Repositorios públicos de Proteómica a gran escala y dirigida	46
1.11.1. PRIDE	46
1.11.2. PeptideAtlas	47
1.12. Formatos de archivos usados en espectrometría de masas y Proteómica	47
1.13. <i>Candida albicans</i> como organismo modelo	51
 OBJETIVOS	 53
Objetivos	55
 II DESARROLLO DE UNA APLICACIÓN WEB PARA RESULTADOS DE IDENTIFICACIONES DE PROTEÓMICA DE <i>Candida albicans</i>	 57
Proteopathogen, a protein database for studying <i>Candida albicans</i> - host interaction	59
Proteopathogen2: A database and web tool to store and display proteomics identification results in the mzIdentML standard	65
 III CREACIÓN DE UN PEPTIDEATLAS DE <i>Candida albicans</i>	 73
A <i>Candida albicans</i> PeptideAtlas	75

Subcellular fractionation and different growing conditions lead to a large increase in the proteome coverage in the <i>Candida albicans PeptideAtlas</i>	83
IV DISCUSIÓN	85
Discusión	87
Desarrollo de una aplicación web para datos de proteómica a gran escala de <i>Candida albicans</i>	88
Desarrollo de un PeptideAtlas <i>Candida albicans</i>	90
V CONCLUSIONES	93
2. Conclusiones	95
Bibliografía	97
Índice alfabético	104
Lista de acrónimos	104

Índice de figuras

1.1. Aumento de complejidad desde el genoma hacia el proteoma	5
1.2. Esquema de un espectrómetro de masas	7
1.3. Ionización MALDI y ESI	9
1.4. Analizadores de masas tipo Cuadrupolo, Trampa Iónica Tridimensional (QIT) y Orbitrap	12
1.5. Espectrometría de masas en Tandem. MS/MS	14
1.6. Nomenclatura de Roepstorff para los fragmentos en MS/MS	16
1.7. Etapas en un experimento de Proteómica <i>Shotgun</i>	20
1.8. Interpretación automática de espectros MS/MS	23
1.9. Estrategia básica de identificación. Selección de péptidos candidatos. Correlación espectro MS/MS - secuencia aminoacídica	26
1.10. Tabla de contingencia. Contraste de hipótesis	31
1.11. Estimación del e-valor	32
1.12. Distribuciones de Espectro Individual y Promedio	34
1.13. Construcción de una base de datos señuelo	35
1.14. Estimación de PeptideProphet de las distribuciones de PSM incorrectos y correctos	39
1.15. Agrupamiento no aleatorio de péptidos en proteínas	42
1.16. Adquisición y reconstrucción de la señal en un experimento SRM . .	46
1.17. Visión general de formatos comúmente usados en cada etapa de un experimento de Proteómica	48

ÍNDICE DE FIGURAS

ÍNDICE DE FIGURAS

1.18. Status de equilibrio entre las células de <i>Candida albicans</i> y células del sistema inmune en las mucosas	52
--	----

Índice de Tablas

INTRODUCCIÓN

Introducción

*Nada en Biología tiene sentido si no es bajo la
luz de la Evolución*

Theodosius Dobzhansky

INTRODUCCIÓN

1.1. Proteómica. Conceptos generales

El concepto de Proteoma fue acuñado originalmente por Marc Wilkins en los años 90 en analogía al concepto de Genoma ([Wasinger et al., 1995](#)). Si el Genoma es la dotación génica de una célula u organismo, el Proteoma es entendido como *el conjunto total de proteínas expresadas por los genes de una célula, tejido u organismo*. Sin embargo, mientras que el Genoma es el mismo en todas las células del organismo, el Proteoma es un concepto más variable. Los genes se expresan en función de las condiciones en que se encuentra la célula, según el orgánulo, el tejido, y estadio del desarrollo entre otros factores. Además existen niveles de complejidad adicional en el curso de información desde el gen a la proteína como el procesamiento alternativo de intrones y las Modificaciones Post-Tradicionales o PTM (*Post-Translational Modification*). Por eso el término Proteoma puede diversificarse, para ajustarse a definiciones mas específicas. Así, podemos hablar del proteoma (o fosfo-proteoma, por ejemplo) de un orgánulo celular, como la mitocondria, en un tejido concreto, en unas condiciones ambientales definidas por los nutrientes disponibles, posiblemente sometida a condiciones de estrés, etc.

Proteómica es, por tanto, el estudio del Proteoma, independientemente del conjunto o subconjunto de proteínas objeto de estudio. Pero además Proteómica se refiere a las tecnologías utilizadas para ello.

El establecimiento de la espectrometría de masas aplicada a moléculas biológicas a finales de los años 80 y el desarrollo de técnicas de separación de proteínas y péptidos como la Electroforesis en Gel de Poliacrilamida, PAGE (*PolyAcrylamide Gel Electrophoresis*) y la Cromatografía Líquida de Alto Rendimiento, HPLC (*High Performance Liquid Chromatography*) permitieron que la Proteómica se consolidara y extendiera como disciplina científica.

La figura 1.1 ilustra como el grado de complejidad biológica desde la unidad de información, es decir, el gen, hasta la unidad funcional, la proteína, aumenta exponencialmente.

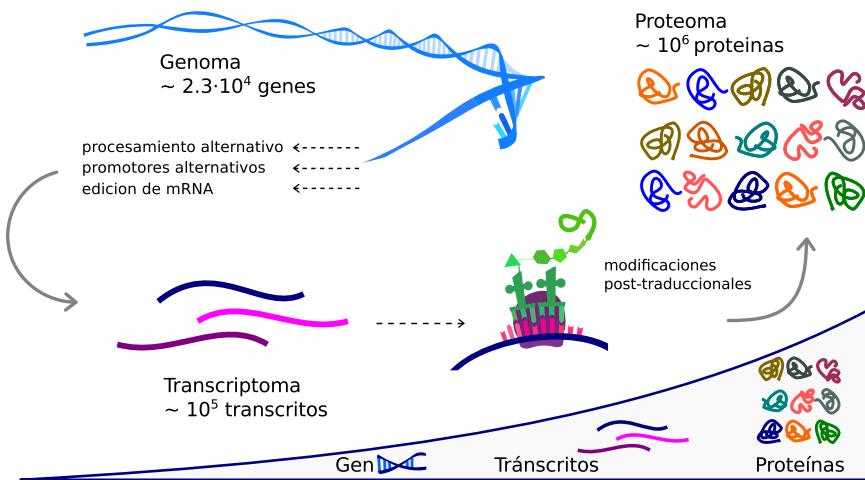


Figura 1.1: Aumento de complejidad desde el genoma hacia el proteoma

1.2. Espectrometría de masas

El desarrollo de las técnicas de ionización suave de macromoléculas biológicas a finales de los años 80, además de valer el Nobel a los químicos John Fenn y Koichi Tanaka, permitió sentar las bases de la Espectrometría de Masas aplicada a la Proteómica. Las técnicas de Ionización por ElectroSpray, ESI (*Electro-Spray Ionization*) (Fenn et al., 1989) y Desorción Suave por Láser, SLD (*Soft Laser Desorption*) (Tanaka et al., 1988) permitieron que las grandes y frágiles moléculas biológicas como las proteínas pudieran ser ionizadas y volatilizadas relativamente intactas para ser posteriormente introducidas en los espectrómetros de masas.

Como ocurre en muchas otras ocasiones en la ciencia, de forma paralela e independientemente habían surgido en distintas partes del mundo ideas muy similares. Así, el desarrollo de SLD que valió el Nobel a K. Tanaka, tuvo un precedente unos años antes. Franz Hillenkamp y Michael Karas en Frankfurt, Alemania (éstos discutiblemente no galardonados) habían ideado una técnica similar que, en este caso, denominaron Ionización mediante Desorción por Láser Asistida por Matriz, MALDI (*Matrix Assisted Laser Desorption Ionization*) (Karas y Hillenkamp, 1988)

INTRODUCCIÓN

Aunque MALDI no fue aplicada a la ionización de proteínas hasta la publicación del trabajo de Tanaka, actualmente éste es el acrónimo que se ha impuesto para referirse a la técnica y es, de hecho, una técnica muy extendida en laboratorios de espectrometría de masas.

1.2.1. Consideraciones sobre unidades empleadas en espectrometría de masas

La unidad fundamental de masa usada en física y química, empleada en la medida de masas atómicas y moleculares, es la llamada *Unidad de Masa Atómica, u*, o **uma** también denominada *Dalton, Da*. La escala de unidades de masa atómica es una escala relativa donde la referencia es el átomo de carbono. El valor de una *u* o un *dalton* se define como la doceava parte de la masa de un átomo neutro de ^{12}C , el isótopo más frecuente de carbono. Así, la masa de un átomo de ^{12}C es de 12 *u*. Y una *u* es aproximadamente equivalente a la masa de un átomo de hidrógeno o la masa de un protón.

$$1\text{Da} = 1\text{u} = 1/12 \cdot \left(\frac{12\text{g } ^{12}\text{C}}{\text{mol } ^{12}\text{C}} \right)$$

$$1\text{Da} = 1\text{u} = 1/12 \cdot \left(\frac{6,0221 \times 10^{23} \text{atmosos } ^{12}\text{C}}{\text{mol } ^{12}\text{C}} \right) \quad (1.1)$$

$$1\text{Da} = 1\text{u} = 1,66054 \times 10^{-24} \text{g/atomos } ^{12}\text{C} = 1,66054 \times 10^{-24} \text{kg/atomos } ^{12}\text{C}$$

Sin embargo los analizadores (espectrómetros) de masas no miden la masa de los analitos ionizados sino la relación entre masa y carga m/z , donde m es la masa molecular del analito y z un múltiplo entero del número de cargas del ion. La unidad empleada para medir esta relación es el *Thomson, Th*. Un thomsom equivale a 1 Da / e. En general, para iones monocargados y solo en ese caso, la masa en Da es equivalente a su valor en thomsons.

Además, en espectrometría de masas es interesante medir la masa exacta de los isótopos de los elementos que componen las moléculas. En este sentido es importante diferenciar entre las masas mono-isotópica y masa promedio

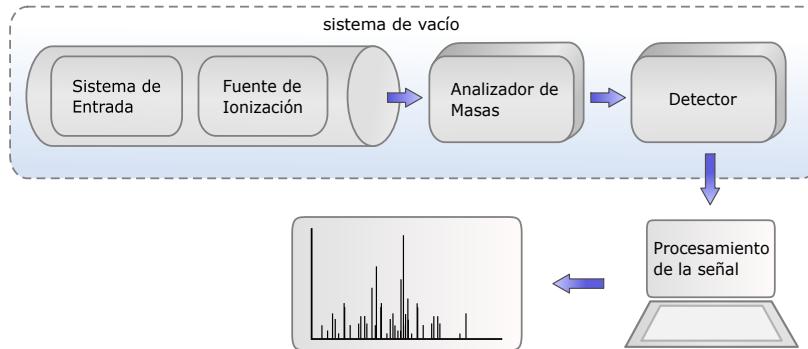


Figura 1.2: Esquema de un espectrómetro de masas

La *masa promedio* es equivalente a una media de las masas atómicas de todos los átomos de los elementos que componen el ion ponderados por abundancia isotópica. Mientras que la *masa mono-isotópica* es aquella en la que se considera que todos los átomos de C se encuentran en su forma ^{12}C .

1.2.2. Componentes de un espectrómetro de masas

Un espectrómetro de masas es, en esencia, una balanza de precisión molecular capaz de medir, hasta un determinado límite de sensibilidad, la masa (en relación a la carga) de moléculas (ionizadas). Consta básicamente de cuatro partes o secciones:

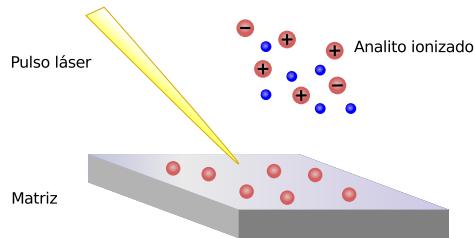
- 1. Sistema de entrada.** Generalmente los espectrómetros de masas se encuentran acoplados con sistemas cromatográficos de alta resolución que permiten que los analitos de una muestra inicialmente muy compleja sean separados e introducidos gradualmente. Este acoplamiento requiere una interfaz, una conexión física y funcional entre el sistema de cromatografía y el espectrómetro, que consiste generalmente en una columna capilar de caudal controlado. En ocasiones, como es en el caso de ESI, el sistema de entrada y la fuente de iones forman parte de un único componente.

INTRODUCCIÓN

2. **Fuente de iones.** Las macromoléculas biológicas, como proteínas y péptidos, no son fácilmente volatilizadas. El desarrollo de las técnicas de ionización suave permitió que péptidos y proteínas ionizados y relativamente intactos pudieran ser introducidos, en fase gaseosa, en un sistema de vacío en los espectrómetros de masas para ser analizados. La ionización ESI y MALDI son las más comunes en Proteómica aunque existen también otros métodos un poco menos utilizados.
 - **MALDI** (Figura 1.3 a) consiste en embeber la muestra en una matriz líquida, que posteriormente se seca, con alta capacidad de absorber luz UV sobre la que inciden pulsos de luz láser UV. Al absorber la energía del láser las moléculas que conforman la matriz son ionizadas por adición de protones que son luego transferidos al analito.
 - En **ESI** (Figura 1.3 b) el analito se encuentra en fase líquida en un solvente orgánico volátil como metanol o acetonitrilo. Esta solución es conducida a través de un capilar sometido a un campo eléctrico de forma que las micro-gotas en el ápice del capilar, una vez que la carga supera un límite, forman un aerosol. Las micro-gotas del aerosol disminuyen su tamaño por evaporación del solvente, reagrupándose en gotas más estables y pequeñas en un proceso reiterativo, hasta el punto en que las moléculas de analito se repelen con la fuerza suficiente para superar la tensión superficial y liberarse (explosión de Coulomb) quedando en suspensión y siendo así introducidos en un sistema vacío hacia el espectrómetro.
- Otros tipos de ionización menos comunes son el Bombardeo Rápido Atómico, FAB (*Fast Atom Bombardment*), la Desorción por Campo Eléctrico, FD (*Field Desorption*), y la Desorción por Plasma, PD (*Plasma Desorption*).

3. **Analizador de masas.** El analizador de masas es la parte del instrumento en la que los iones se separan en base su relación entre la masa y carga (m/z). Es el elemento que se usa generalmente para definir el tipo de instrumento. Existen varios tipos que pueden combinarse en los llamados espec-

a) MALDI



b) ESI

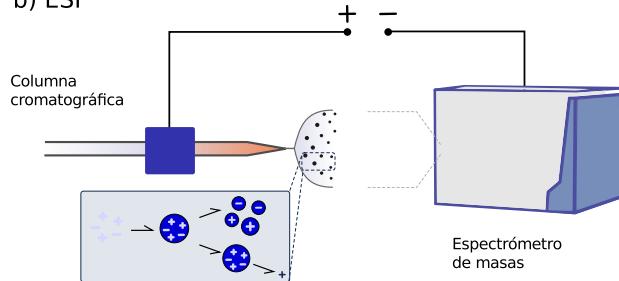


Figura 1.3: Ionización MALDI y ESI

trómetros híbridos. Así, un analizador de tipo cuadrupolo puede encontrarse acoplado con un analizador de *tiempo de vuelo* o una *trampa iónica* para formar un Cuadrupolo-Tiempo de Vuelo, QTOF (*Quadrupole-Time Of Flight*) o Cuadrupolo-Trampa Iónica, QTrap (*Quadrupole-Ion Trap*) respectivamente. Algunos de los analizadores de masas más comunes son los siguientes:

- Los **Analizadores de sector** (magnético o eléctrico) aceleran los iones de analito que al atravesar el sector son sometidos a un campo con fuerza ortogonal a la trayectoria del ion lo que provoca que se desvíen en función de su relación m/z .
- Los **Analizadores de Tiempo de Vuelo**, TOF (*Time Of Flight*) Este tipo de analizador usa un campo eléctrico para acelerar los iones de analito. La separación se produce por la diferencia en el tiempo que éstos invierten

INTRODUCCIÓN

en recorrer una distancia en el vacío en el interior del analizador, el llamado *Tiempo de Vuelo*. La aceleración y por tanto el Tiempo de Vuelo es una función de la relación m/z de los iones que impactan en el detector a diferentes tiempos. Por su carácter dependiente de la dimensión tiempo, los analizadores TOF se usan generalmente en combinación con ionización MALDI, que introduce iones en el analizador en pulsos de láser.

- **Cuadropolos.** Los analizadores de tipo Cuadropolo (Figura 1.4 a) reciben su nombre porque constan de cuatro varillas metálicas enfrentadas en pares llamados polos con cargas opuestas. Sobre estos pares, además del potencial eléctrico de corriente continua, se aplica también una corriente alterna de radiofrecuencia. Esta conformación permite crear un campo eléctrico oscilante controlado que estabiliza (o desestabiliza) selectivamente los iones que pasan a través y de esta forma solo los iones con ciertos valores m/z podrán llegar a impactar en el detector mientras que el resto son desviados y filtrados.
- Las **Trampas Iónicas** funcionan bajo el mismo principio físico que los cuadropolos, pero la conformación en forma de cámara de las trampas iónicas permite confinar y acumular los iones que luego son liberados selectivamente.

Las Trampas Iónicas Tridimensionales, QIT (*Quadrupole Ion Trap*) (Figura 1.4 c) constan de dos electrodos metálicos de sección hiperbólica enfrentados y un electrodo toroidal que conforman una cámara donde se acumulan los iones de analito. En el interior los iones orbitan en el vacío. El ajuste de la radiofrecuencia permite filtrar selectivamente los iones, estabilizando aquellos con determinados valores m/z y desestabilizando el resto, que colisionan con el electrodo y no llegan al detector.

Las Trampas Iónicas Lineales, LTQ (*Linear Trap Quadrupole*) o LIT (*Linear Ion Trap*) consisten en un sistema de cuadropolo, que sitúa los iones en un eje radial, y dos electrodos terminales, uno en cada extremo, que confina los iones longitudinalmente.

Orbitrap (Figura 1.4 b) es un tipo de trampa iónica, relativamente reciente, desarrollado a finales de los años 90 del s.XX (Makarov, 2000). Consiste en un electrodo en un eje interno rodeado por un electrodo externo cilíndrico. Los iones son introducidos tangencialmente desde la fuente de ionización y, al ajustar la diferencia de potencial, son atrapados en órbitas elípticas longitudinales, en las que la atracción hacia el eje interno es compensada por la fuerza centrífuga. La relación m/z se determina a partir de la frecuencia angular de la oscilación de los iones en torno al electrodo longitudinal interno.

Los analizadores de tipo Resonancia Iónica Ciclotrónica - Transformada de Fourier, FTICR (*Fourier Transform Ion Cyclotron Resonance*) se basan en confinar los iones en una celda ICR donde un campo magnético homogéneo somete los iones a seguir una trayectoria circular con una frecuencia de rotación característica de cada relación m/z y del valor del campo. Al aplicar un campo eléctrico de igual frecuencia a la frecuencia de rotación, la partícula es excitada para seguir una trayectoria más larga aumentando el radio de giro provocando que los iones alcancen las placas de detección. La señal formada por la mezcla de las frecuencias de todos los iones es entonces deconvolucionada mediante la transformada de Fourier que permite la detección simultánea de todas las frecuencias.

4. **Detector** El detector es elemento final de un espectrómetro de masas. Registra la corriente producida por el haz de iones que incide sobre él convirtiéndola en una señal eléctrica medible. Los detectores más utilizados en espectrometría de masas son los *multiplicadores de electrones*. Este tipo de detector utiliza la energía cinética de los iones que inciden sobre una placa que tiene su superficie recubierta por óxidos de tierras raras; al chocar los iones contra la placa, ésta emite una corriente de electrones que son acelerados hacia una segunda placa, de la que vuelven a arrancar electrones que son acelerados hacia una tercera placa y así sucesivamente para conseguir la amplificación de la señal.

INTRODUCCIÓN

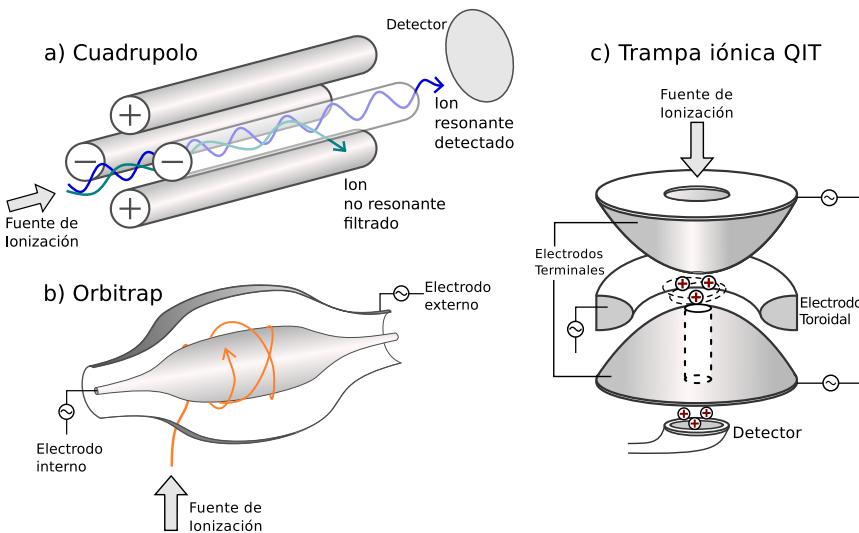


Figura 1.4: Analizadores de masas tipo Cuadrupolo, Trampa Iónica Tridimensional (QIT) y Orbitrap

La *sensibilidad*, *resolución*, *precisión* y *exactitud* son parámetros importantes en espectrometría de masas ya que determinan notablemente la cantidad y calidad de información del espectro generado, lo que a su vez, es esencial para identificar el péptido que origina el espectro.

La *sensibilidad* de un espectrómetro de masas es la capacidad para detectar masas muy pequeñas. Puede llegar a ser de hasta unas pocas partes por millón (ppm) en el caso de instrumentos de alta precisión como LTQ-Orbitrap, pero requiere un ajuste óptimo de múltiples parámetros como la calibración del instrumento o la temperatura entre otros.

La *resolución* es la capacidad para discernir señales que realmente corresponden a diferentes iones dentro de una ventana o margen de valores m/z . Esto es esencial para evitar la co-fragmentación, es decir, obtener fragmentos de iones precursores diferentes con valores m/z similares, o cuando se requiere conocer la distribución isotópica de un ion. La resolución se define como la diferencia entre las

masas de dos picos adyacentes que están resueltos

$$R = \frac{M}{\Delta M} \quad (1.2)$$

donde M es el valor entero más próximo de masa del primer pico y ΔM es el incremento de m/z a una determinada altura del pico. Frecuentemente se usa un 50 % de altura del pico, el parámetro en ese caso es la Altura de Pico a Media Altura, FWHM (*Full Width at Half Mass*). La resolución es un parámetro específico del espectrómetro de masas. Otros parámetros, relacionados aunque no específicos del instrumento son la exactitud y la precisión.

La *exactitud* de la medida de la masa molecular es la diferencia entre el valor m/z obtenido para un ion y su valor real. Se expresa generalmente en partes por millón (ppm) y representa el error del valor m/z obtenido con respecto al valor verdadero.

La *precisión* es una medida de la dispersión de una serie de mediciones obtenidas para un analito determinado (en condiciones experimentales equiparables) con respecto a un valor promedio de referencia. El parámetro que se emplea generalmente como medida de precisión es la desviación estándar σ , equivalente a la raíz cuadrada de la varianza.

1.2.3. Espectrometría de masas en Tandem. MS/MS

Los péptidos, separados en el espectrómetro de masas en base a su relación m/z , generan señales cuyas intensidades son registradas en el detector e interpretadas como un espectro. El objetivo básico en Proteómica consiste en la elección del mejor péptido candidato, y por extensión la inferencia de la proteína original, responsable de los espectros obtenidos. La aproximación general, en esencia, consiste en estimar el grado de similitud entre los valores m/z empíricos obtenidos en el espectro y los valores m/z calculados que teóricamente se producen a partir de una digestión predicha computacionalmente de las secuencias en una base de datos de referencia.

En ocasiones, cuando la proteína original se encuentra relativamente aislada, el espectro que generan los péptidos que se detectan en el instrumento es suficiente-

INTRODUCCIÓN

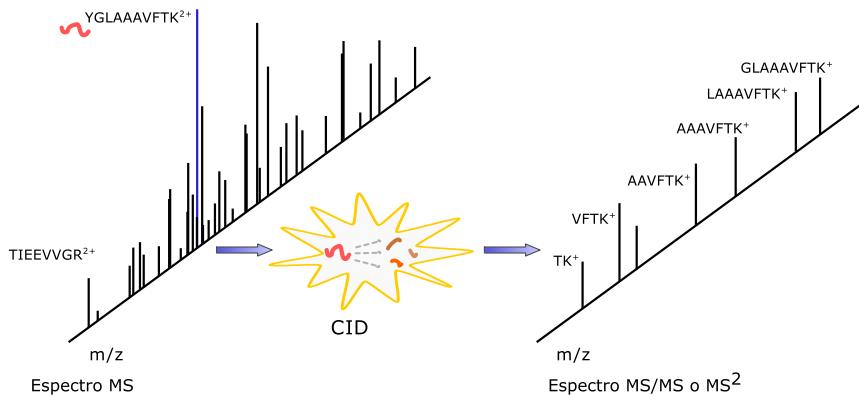


Figura 1.5: Espectrometría de masas en Tándem. MS/MS

mente informativo y específico de la proteína original y ésta puede ser identificada. Este es el principio de la técnica conocida como Huella Peptídica, descrita en la sección 1.4.1. Sin embargo, esta técnica requiere que la proteína se encuentre aislada y el rendimiento que ofrece es, por tanto, limitado.

En la Espectrometría de Masas en Tándem, MS/MS (*Tandem Mass Spectrometry*) o MS², los péptidos, una vez ionizados y dentro del espectrómetro, son sometidos a una fragmentación adicional. (Figura 1.5). Los péptidos se fragmentan, generando iones más pequeños lo que hace que el patrón de fragmentación sea más específico de la secuencia original. Esto aumenta el poder de resolución del análisis, ya que permite distinguir péptidos que, intactos, tienen masas muy similares, pero cuyos patrones de fragmentación MS/MS son diferentes. Esto posibilita, además, partir de muestras con mezclas de proteínas más complejas incrementando así el rendimiento del experimento.

El proceso que conduce a la adquisición de espectros conlleva varias etapas. En primer lugar el instrumento escanea todos los péptidos ionizados introducidos en el espectrómetro y registra los llamados espectros MS¹, valores *m/z* y sus correspondientes intensidades para cada ion. A continuación, en función de la intensidad registrada en MS¹, se seleccionan y aislan algunos de estos iones -*precursores*- para ser fragmentados en péptidos más pequeños -*fragmentos*- en el interior del

espectrómetro. El espectro MS² adquirido o espectro de fragmentación, registra los valores m/z e intensidades de los fragmentos de cada uno de los péptidos precursores aislados y fragmentados. El patrón de fragmentación codificado en los espectros MS² contiene la información necesaria para deducir la secuencia aminoacídica del péptido que lo origina.

En algunos análisis puede ser necesario realizar fragmentaciones adicionales que permitan un mayor aún poder de resolución. Estos análisis se conocen como MSⁿ, donde n es el número de fragmentaciones y etapas de análisis de masas consiguientes.

A este método de adquisición de espectros de fragmentación, en el que los péptidos precursores son seleccionados para ser fragmentados en base las intensidades en el espectro MS¹ se denomina por eso Adquisición Dependiente de Datos, DDA (*Data Dependent Acquisition*)

Los fragmentos originados a partir del péptido, según la nomenclatura de Roepstorff ([Roepstorff y Fohlman, 1984](#)), se clasifican, en función del punto donde se produce la ruptura, en las denominadas series x , y y z si la carga del ion permanece en el extremo carboxilo-terminal y las series a , b y c si la carga permanece en el extremo amino-terminal. Además se añade un sub-índice que indica el número de residuos en el fragmento (Figura 1.6). Generalmente, los iones más abundantes e informativos son los b - e y -, generados por la fragmentación en el enlace peptídico entre aminoácidos, el punto de menor energía de la estructura. En analizadores tipo cuadrupolo o QTOF predominan los iones y -, mientras que en las trampas iónicas se generan igualmente b - e y - ([Steen y Mann, 2004](#))

Técnicas de fragmentación

La Disociación Inducida por Colisión, CID (*Collision Induced Dissociation*) es uno de los métodos de fragmentación más frecuentemente utilizados en espectrometría de masas para Proteómica. Consiste en hacer colisionar a las moléculas de analito con átomos o moléculas de gases nobles, químicamente inertes. Argón o Xenón son generalmente usados en triples cuadrupolos y Helio en las trampas iónicas. La colisión provoca que parte de la energía cinética del ion sea transformada en energía vibracional lo que provoca la ruptura del esqueleto peptídico. La

INTRODUCCIÓN

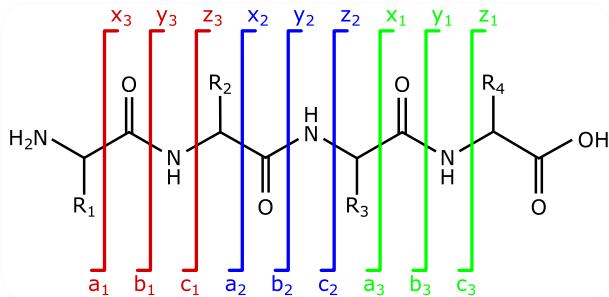


Figura 1.6: Nomenclatura de Roepstorff para los fragmentos en MS/MS

fragmentación tipo CID genera una elevada proporción de iones de las series *b*- e *y*-

Otros tipos de fragmentación son la Disociación por Transferencia de Electrones, ETD (*Electron Transfer Dissociation*), (Syka et al., 2004), en la que los iones de analito con carga positiva interaccionan con aniones que les transfieren un electrón produciendo una fragmentación con presencia de iones *c*- y *z*-; la Disociación por Captura de Electrones, ECD (*Electron Capture Dissociation*) (Zubarev et al., 1998), consistente la interacción del analito con electrones suministrados directamente; y, por último, la Disociación por Colisión de Alta Energía, HCD (*Higher Energy Collision Dissociation*) usado en analizadores tipo Orbitrap (Olsen et al., 2007), y que al igual que CID genera iones *b*- e *y*-, si bien, debido a la mayor energía de activación, los iones *b*- sufren fragmentaciones adicionales generando iones *a*- y otras especies de menor tamaño.

1.3. Digestión de proteínas en péptidos

Tras la obtención de una muestra de proteínas, ya sea una mezcla compleja o una proteína más o menos aislada y purificada, el primer paso en un experimento de Proteómica consiste en someter a las proteínas a la acción de una enzima proteasa que corta en puntos específicos de la secuencia y que las digiere en un conjunto de péptidos. Sin embargo, sabiendo que ciertos espectrómetros de masas tienen la capacidad de medir masas de proteínas intactas, podemos preguntarnos:

¿por qué hacer una digestión que aumenta el grado de complejidad de la muestra y que supone el problema añadido de la inferencia de la proteína originaria a partir de sus péptidos constituyentes? o dicho de otra manera ¿es necesario el paso intermedio de digestión en péptidos para luego inferir las proteínas originales?

La respuesta a estas preguntas, revisada en (Steen y Mann, 2004), tiene que ver, sobre todo, con limitaciones técnicas. Las proteínas intactas pueden ser difíciles de manipular, algunas, como las proteínas de membrana son insolubles en condiciones en que otras sí lo son. Muchos detergentes comúnmente usados interfieren en MS ya que son fácilmente ionizables y se encuentran en gran cantidad en proporción a las proteínas (Hatt et al., 1997). Además la sensibilidad de los espectrómetros es menor para proteínas intactas que para péptidos.

La digestión consiste en la rotura de proteínas en péptidos por acción de una enzima proteolítica. Tradicionalmente se ha utilizado para esto *tripsina*, que rompe la secuencia aminoacídica a continuación, en el lado carboxilo-, de Arginina (R) o Lisina (K) a menos que exista una Prolina (P) adyacente. Los péptidos generados por acción de la tripsina, llamados péptidos *trípticos*, tienen un tamaño adecuado, dada la frecuencia media de R y K, para el análisis por espectrometría de masas lo que explica la popularidad de esta proteasa.

También es posible la utilización de otras proteasas siempre que se conozca su patrón de corte. Es de hecho una aproximación inevitable para aquellos casos en que la tripsina no sea útil, por ejemplo, debido a una baja frecuencia de R y K que no generen péptidos del tamaño adecuado.

1.4. Proteómica en gel

La separación de proteínas por Electroforesis en Gel de Poliacrilamida, PAGE, es una técnica, o serie de técnicas con variantes, que consiste en separar proteínas presentes en una muestra inicial en base a propiedades fisico-químicas diferenciadoras como su carga, tamaño y/o su punto isoeléctrico. En función del número de estas propiedades que se aprovechan para separar, en mayor o menor grado, las proteínas de una muestra se distinguen básicamente dos tipos de PAGE

En la Electroforesis en Gel de Poliacrilamida Monodimensional, 1D-PAGE (*Mo-*

INTRODUCCIÓN

nodimensional PoliAcrylamide Gel Electrophoresis), las proteínas se separan en función de su peso molecular, las más pequeñas se separan más en el gel. En la Electroforesis en Gel de PoliAcrilamida Bidimensional, 2D-PAGE (*Bidimensional PoliAcrylamide Gel Electrophoresis*), las proteínas se separan en una primera dimensión (sobre una tira con un gradiente de pH inmobilizado) en función de su punto isoeléctrico para posteriormente, aplicar la segunda dimensión, equivalente a 1DPAGE.

Otra clasificación posible de las técnicas PAGE puede establecerse en función de si se usan condiciones desnaturalizantes o no. Entre las técnicas que usan geles desnaturalizantes probablemente la más empleada sea la Electroforesis en Gel de PoliAcrilamida con DodecilSulfato Sódico, SDS-PAGE (*Sodium Dodecyl Sulfate PoliAcrylamide Gel Electrophoresis*). Y entre las que usan condiciones nativas o no desnaturalizantes, la llamada Blue Native, empleada para estudiar proteínas agrupadas en complejos.

La Proteómica en gel ha sido (y continúa siendo) una técnica muy empleada en laboratorios de todo el mundo. Tiene algunas limitaciones, como el hecho de que proteínas de bajo peso molecular no son fácilmente observables, o que el número de proteínas identificables a partir de un gel difícilmente pueda superar el millar. Sin embargo, este tipo de estudios sigue teniendo un nicho en la Proteómica actual ([Rogowska-Wrzesinska et al., 2013](#)). Notablemente, permite la visualización, identificación y cuantificación de proteínas intactas. La particular capacidad de la Proteómica en gel para separar proteínas con pequeños cambios en sus puntos isoeléctricos, *pI*, permite discernir entre isoformas de proteínas, o versiones de la misma proteína con PTM, lo que difícilmente se puede conseguir con otro tipo de aproximaciones.

1.4.1. Huella Peptídica

La Huella Peptídica de una proteína se refiere al hecho de que el patrón de fragmentación de una proteína en los péptidos que la constituyen utilizando una enzima proteolítica determinada, es muy específico de la proteína originaria (siempre y cuando se conozca el patrón de corte de la enzima, como es el caso de la tripsina) de forma que el espectro que generan puede ser utilizado para identificar-

la. Sin embargo, a pesar de esta especificidad, la enorme variedad de proteínas implica una mayor aún variedad de posibles péptidos generados a partir de ellas que pueden tener masas muy similares. Por ese motivo, esta técnica requiere que la proteína se encuentre previamente aislada, generalmente a partir de una *mancha o spot* proteico de 2D-PAGE

Generalmente la técnica de la Huella Peptídica, conocida también por el acrónimo PMF, PMF (*Peptide Mass Fingerprint*), se lleva a cabo por espectrometría de masas MALDI-TOF(ToF). Esto significa que, una vez obtenidos los péptidos correspondientes a la proteína del *spot*, éstos se sitúan en una matriz MALDI, donde son ionizados e introducidos en un analizador TOF. Una vez obtenido el espectro patrón de masas peptídicas, el proceso de análisis consiguiente es similar al que se hace en la Proteómica a gran escala o *shotgun*. Como se describe en las secciones siguientes, la identificación del péptido responsable del espectro se realiza utilizando un motor de búsqueda, que compara los valores de *m/z* del espectro obtenidos empíricamente con los valores de *m/z* calculados a partir de las secuencias de péptidos tripticos teóricos.

1.5. Proteómica a gran escala

La Proteómica a gran escala, conocida generalmente por el término inglés *shotgun*, es la técnica de elección para la mayoría de estudios proteómicos enfocados a obtener un alto rendimiento. El nombre *shotgun* proviene de una analogía con las técnicas clásicas de secuenciación genómica donde el ADN es fragmentado en secuencias más pequeñas que posteriormente son ensambladas. En la Proteómica *shotgun* las proteínas son fragmentadas en péptidos a partir de los cuales se infiere finalmente la proteína original. Implica varios pasos descritos en las siguientes secciones.

1.5.1. Separación de péptidos y proteínas sin gel

A diferencia de la técnica de la Huella Peptídica donde cada proteína se encuentra relativamente aislada, en la Proteómica de alto rendimiento o *shotgun*, puesto que el objetivo es identificar el máximo número de proteínas en un solo experimento.

INTRODUCCIÓN

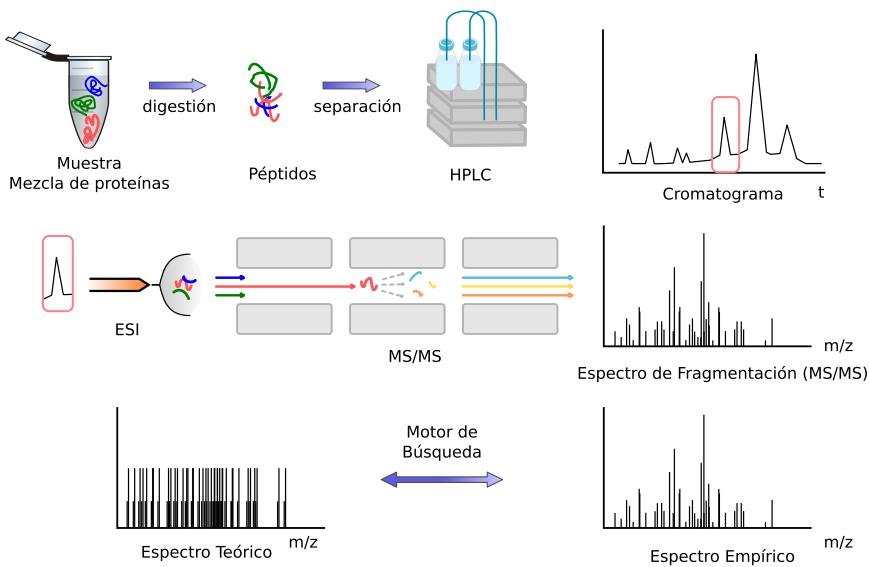


Figura 1.7: Etapas en un experimento de Proteómica *Shotgun*

to, se parte de una muestra más compleja. Esto es importante porque, sabiendo que a partir de cada proteína se generan múltiples péptidos (trípticos), el grado de complejidad de la muestra aumenta enormemente tras la digestión. Por este motivo, para evitar que la mezcla de péptidos sea demasiado compleja para la resolución en el análisis MS, previamente a la introducción de los péptidos en el espectrómetro, se realiza una cromatografía que permite separar los péptidos para que sean ionizados y lleguen al analizador de masas gradualmente.

Opcionalmente esta separación puede comenzar a nivel de proteína por electroforesis en un gel 1D-PAGE , o incluso a un nivel estructural superior, por ejemplo, por fraccionamientos sub-celulares correspondientes a distintos orgánulos.

Cromatografía Líquida de Alto Rendimiento. HPLC

Pero el fraccionamiento más importante se hace a nivel de péptido, trás la digestión de las proteínas, por HPLC. El funcionamiento básico general en HPLC con-

siste en hacer pasar la muestra a través de una fase estacionaria en el interior de una columna mediante el bombeo a alta presión de una fase móvil. De esta forma los componentes de la muestra se retrasan diferencialmente en función de sus interacciones químicas con la fase estacionaria a medida que atraviesan la columna. La fase móvil suele ser una combinación, en proporciones variables, de un componente acuoso al que se añade un ácido (trifluoroacético o fórmico) y un solvente orgánico (comúnmente acetonitrilo o metanol). Esta proporción en la composición de la fase móvil puede ser constante (cromatografía isocrática) o variable, en gradiente de elución. En un gradiente típico, al aumentar la proporción del solvente orgánico, los analitos de la muestra irán progresivamente teniendo mayor afinidad por la fase móvil y se separan de la fase estacionaria. El *Tiempo de Retención* o *Tiempo de Elución* es el tiempo que necesita un analito para atravesar la columna. Siempre que las condiciones cromatográficas permanezcan invariables el tiempo de retención de un analito es una característica identificativa.

El tipo más común de cromatografía usada en experimentos de Proteómica es la que se conoce como Cromatografía Líquida de Alto Rendimiento en Fase Reversa, RP-HPLC (*Reverse Phase High Performance Liquid Chromatography*) En ella los analitos de la muestra se separan en base a su carácter hidrofóbico. La fase estacionaria, apolar, está compuesta por unas micro-esferas de sílice cubiertas de cadenas alquilo con 18 átomos de C (C18). Un gradiente en que el solvente orgánico aumente gradualmente y en proporción inversa al componente acuoso, provoca que los analitos más polares eluyan primero integrados en la fracción acuosa cuando hay una mayor proporción de ésta, mientras que los más hidrofóbicos son retenidos durante más tiempo.

A continuación, una vez producida la separación cromatográfica, los péptidos son ionizados e introducidos en el espectrómetro de masas. En ocasiones, como el caso de la ionización ESI, el sistema de entrada y la fuente de iones forman parte de un único componente que se encuentra acoplado (*on-line*) al espectrómetro de masas.

Tras la adquisición experimental de los espectros, el paso siguiente en un experimento de Proteómica *shotgun* implica el análisis computacional de esos espectros cuyo objetivo final es la obtención de una lista de proteínas que presumiblemente

INTRODUCCIÓN

se encuentran en la muestra. Este análisis computacional, a su vez, consta de varios procesos secuenciales, principalmente la **asignación de secuencias peptídicas** a cada espectro (sección 1.6), la **inferencia de las proteínas** a partir de esos péptidos (sección 1.8) y una **evaluación estadística** que aporta medidas de fiabilidad a la **identificación** (sección 1.7).

1.6. Asignación Péptido-Espectro

En un experimento típico de Proteómica *shotgun* pueden generarse miles de espectros por hora. La interpretación manual, por lo tanto no es una opción práctica. Diversas aproximaciones computacionales y herramientas de *software* se han desarrollado para facilitar esta tarea de asignación de secuencias peptídicas a los espectros MS/MS. A cada una de estas parejas péptido-espectro se les denomina generalmente **Asignación Péptido-Espectro, PSM** (*Peptide-Spectrum Match*).

Las estrategias utilizadas para la obtención de una lista de PSM básicamente pueden clasificarse en tres tipos. La más extendida es la **búsqueda utilizando bases de datos de secuencias**, consistente en establecer una correlación entre el espectro MS/MS obtenido empíricamente y espectros teóricos predichos a partir de secuencias. Otra estrategia, usada en casos en que el genoma del organismo objeto de estudio no está (o sólo parcialmente) secuenciado, es la **secuenciación *de novo*** en la que la secuencia se infiere directamente del espectro sin ayuda de una base de datos de referencia. El tercer tipo de aproximación, la **búsqueda basada en bibliotecas de espectros**, requiere una recopilación, lo más extensa posible, de espectros MS/MS adquiridos previamente y ya asignados a péptidos, que son comparados directamente con los espectros empíricos adquiridos.

1.6.1. Búsqueda en bases de datos de secuencias

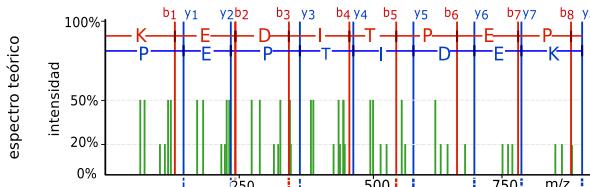
La búsqueda utilizando bases de datos de secuencias es el principal y más extendido método de asignación de una secuencia peptídica a un espectro MS/MS (figura 1.8). Existen una gran variedad de herramientas computacionales llamadas motores de búsqueda diseñadas para realizar esta tarea.

Los motores de búsqueda son un tipo de programas informáticos a los que

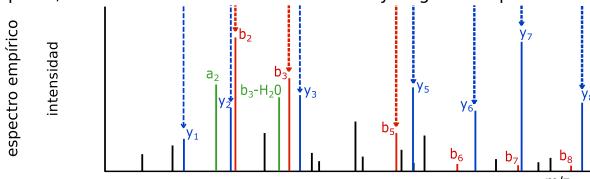
Interpretación automática usando espectros teóricos. Principios básicos

- 1 - Dado el conocimiento previo sobre:
 · Secuencias en BBDD de proteínas
 · Probabilidad de que cada tipo de ión se encuentre en un espectro MS/MS:
- | | |
|-----------------------|---|
| Iones b : 100% | Iones b , y + 2H : 50% |
| Iones a : 20% | Iones b , y -NH ₃ : 20% |
| Iones y : 100% | Iones b , y -H ₂ O : 20% |

- 2 - Se puede crear un espectro teórico para una lista de péptidos candidatos elaborada conociendo:
 · Enzima proteolítica utilizada
 · Rango de tolerancia de masas
 · etc
- Para cada espectro empírico:
 $\{c \in P : |M_{\text{teórica}} - M_{\text{empírica}}| < \text{Tol}\}$
- c:**péptido candidato **P:** péptidos (trípticos) en BD
P E P T I D E K **Tol:** tolerancia



- 3 - Y para cada uno de los espectros teóricos obtenidos comparar con el espectro empírico, obtener todos los iones coincidentes y asignar una puntuación



- 4 - Por último, en base a la puntuación, se puede obtener una lista de los mejores péptidos candidatos que posiblemente han originado el espectro

Figura 1.8: Interpretación automática de espectros MS/MS

se les suministra como entrada datos correspondientes a una lista de espectros MS/MS empíricos y una serie de parámetros que tener en cuenta para restringir la búsqueda. El programa compara estos espectros reales registrados con espectros teóricos que es posible obtener gracias a diversas fuentes de información disponible como el patrón de corte de la enzima proteolítica utilizada, los valores *m/z* de los fragmentos que se producirían a partir de los péptidos, la frecuencia estimada de cada tipo de fragmento y las secuencias de las proteínas en una base de datos de referencia. En el proceso, el espacio de búsqueda se acota mediante la selección de una lista de péptidos posibles, (candidatos que cumplen unos criterios determinados), que posiblemente han generado el espectro MS/MS y a continuación se

INTRODUCCIÓN

ordenan utilizando una puntuación función del grado de similitud entre el espectro empírico y el teórico.

Elaboración de una lista de péptidos candidatos

Para reducir el espacio de búsqueda entre todos los posibles péptidos candidatos que explican el espectro MS/MS y así reducir el coste computacional, el motor de búsqueda requiere, como estrategia heurística, una serie de parámetros proporcionados por el usuario. Éstos, básicamente, reflejan conocimiento previo sobre el experimento y pueden ser entendidos como información auxiliar para facilitar la distinción entre identificaciones auténticas o reales e identificaciones falsas. Los más importantes de estos parámetros son la enzima utilizada y el rango de masas en el que debe encontrarse el ion precursor.

- La selección de la enzima proteolítica utilizada limita la digestión predicha computacionalmente a aquellos péptidos que cumplan el patrón de corte conocido, filtrando el resto de posibles péptidos. Con esto se reduce enormemente el número de comparaciones que el motor de búsqueda debe realizar y, por tanto, el tiempo empleado para ello. Sin embargo, al restringir el tipo de enzima, se imposibilita la identificación de péptidos con rupturas inespecíficas (por ejemplo el procesamiento post- traduccional que provoca la liberación del péptido señal o por proteasas contaminantes presentes en la muestra)
- El establecimiento de un rango o ventana de tolerancia de masas, tanto a nivel de péptido precursor como a nivel de fragmentos, permite excluir aquellos péptidos y fragmentos que se encuentren fuera de dicho rango. Solo los espectros teóricos de aquellos péptidos que cumplen este requisito son comparados con el espectro empírico y puntuados en base a su similitud. La elección de esta tolerancia depende del tipo de espectrómetro utilizado, así, para equipos de alta resolución tipo Orbitrap o FTICR se puede ajustar a valores inferiores a 1 Da.

Otros parámetros que se proporcionan al *software* y que afectan notablemente a la creación de la lista de candidatos y por tanto también al coste computacional

son, la selección de masa mono-isotópica o masa promedio (es decir, considerar que todos los átomos de C se encuentran en su forma ^{12}C o bien que exista una proporción variable de isótopos ^{13}C); el número de puntos de corte no efectuados permitidos dentro de la secuencia del precursor; la existencia de modificaciones post-traduccionales y otras modificaciones permitidas (variables o fijas) que ocurren en el proceso experimental; y la selección del tipo de iones fragmento a buscar.

El establecimiento de estos valores tiene consecuencias muy notables en los resultados de identificación de péptidos y en consecuencia de proteínas. Por ejemplo, restringir a un rango de tolerancia muy pequeño el valor posible de masa del precursor, aunque puede ser útil para obtener espectros de gran calidad en instrumentos muy sensibles, puede dejar fuera secuencias válidas.

Una aproximación sensata puede consistir en (disponiendo de recursos computacionales suficientes) realizar una búsqueda muy abierta y posteriormente refinarla. En ocasiones se puede hacer una búsqueda con una ventana de tolerancia amplia para la masa del precursor y posteriormente emplear ese parámetro en el post-procesamiento (PeptideProphet, sección 1.7.4) de forma que se puede obtener un mayor rendimiento (en términos de identificaciones para una cierta tasa de error) comparado con una búsqueda para el mismo set de espectros con una ventana más restrictiva ([Ding et al., 2008](#); [Nesvizhskii, 2010](#)).

Motores de búsqueda. Funciones de puntuación

Los motores de búsqueda se encargan de asignar a cada espectro empírico obtenido un péptido, el mejor candidato de una lista de los posibles péptidos que han generado ese espectro, con una cierta medida de puntuación función del grado de similitud entre espectro empírico y teórico. La estrategia general consiste en realizar una digestión teórica de las secuencias de proteínas de referencia, teniendo en cuenta los parámetros especificados. Así, para cada espectro observado, el motor de búsqueda recorre las secuencias en una base de datos (un archivo FASTA) seleccionando aquellos péptidos con valores m/z similares al del ion precursor en el espectro empírico y que se encuentran dentro del rango de tolerancia permitido obteniendo un espectro teórico para cada uno. A continuación se establece el grado de similitud de cada espectro adquirido con los espectros teóricos de cada uno de

INTRODUCCIÓN



Figura 1.9: Estrategia básica de identificación. Selección de péptidos candidatos. Correlación espectro MS/MS - secuencia aminoacídica

los péptidos candidatos, es decir, se evalúa la calidad de cada PSM.

Los motores de búsqueda realizan esta comparación de diferentes maneras, usando distintas funciones de puntuación. Algunos incluso calculan más de un tipo de puntuación. Existe una gran variedad de estrategias de puntuación descritas profusamente en la bibliografía, basadas en funciones de correlación entre espectros, basadas en contar el número de fragmentos compartidos, en alineamiento de espectros o en el uso de reglas derivadas más complejas.

SEQUEST (Eng et al., 1994) fue la primera herramienta descrita para correlacionar espectros MS/MS con secuencias de aminoácidos y actualmente sigue siendo uno de los programas más utilizados. Para cada espectro adquirido, SEQUEST calcula de manera independiente la puntuación de correlación (*cross-correlation Score, Xcorr*) para todos los candidatos con los que es comparado. En primer lugar se crea un espectro empírico procesado (espectro X) en el que los picos de baja intensidad son eliminados y el resto de valores *m/z* son redondeados al valor entero más próximo. Para cada candidato se crea un espectro teórico (espectro Y) usando unas reglas de fragmentación simplificadas. Entonces el valor *Xcorr* es calculado como una función de correlación *Corr(t)* (el producto entre los vectores X e Y, con

Y desplazado t unidades de masa respecto a X a lo largo del eje m/z)

$$\begin{aligned} \text{Corr}(t) &= \sum_i x_i y_{i+t} \\ \text{Xcorr} &= \text{Corr}(0) - \langle \text{Corr}(t) \rangle_i \end{aligned} \quad (1.3)$$

Básicamente, Xcorr contabiliza el número de fragmentos coincidentes entre el espectro empírico (procesado) y el espectro teórico permitiendo pequeños desplazamientos. Además la puntuación se corrige teniendo en cuenta una estimación del número de coincidencias entre picos aleatorias. SEQUEST también proporciona un valor de puntuación adicional, ΔCn , que indica la diferencia entre el valor Xcorr del mejor candidato y el del segundo mejor candidato. Ambos valores son por tanto indicativos de la calidad de cada PSM que será mejor cuanto más altas sean ambas puntuaciones.

Una adaptación no comercial, de código abierto, del algoritmo original de SEQUEST es el motor de búsqueda Comet ([Eng et al., 2013](#)), que introduce la novedad de permitir paralelizar el proceso de búsqueda para poder ser ejecutado en procesadores multi-núcleo.

Otro motor de búsqueda frecuentemente utilizado es *X!Tandem* ([Fenyö y Beavis, 2003](#)), que calcula una puntuación llamada *hyperscore*. Ésta también se basa en contar el número de picos compartidos entre los espectros teórico y empírico, pero en este caso, en la versión original del software, se tiene en cuenta si los iones coincidentes pertenecen a las series *b*- e *y*-.

$$\begin{aligned} \text{by-Score} &= \sum \text{Intensidad picos coincidentes } b \text{ e } y - \\ \text{hyperscore} &= (\text{by-Score}) \cdot N_y! \cdot N_b! \end{aligned} \quad (1.4)$$

Opcionalmente, *X! Tandem* puede ser modificado con la puntuación-k ([MacLean et al., 2006](#)), un producto escalar similar al implementado en Comet con una manipulación previa de las intensidades de los iones de los espectros teóricos candidatos.

El motor de búsqueda OMSSA (*Open Mass Spectrometry Search Algorithm*) ([Geer et al.](#)), al igual que *X!Tandem*, es de código abierto. En OMSSA, la puntuación

INTRODUCCIÓN

es un reflejo del número de coincidencias entre iones del espectro experimental y el teórico sin tener en cuenta si los iones son de tipo *b*- o *y*.

Tanto X!Tandem como OMSSA proporcionan una medida adicional además de su propio método de puntuación, el *e-valor*, que da idea de la calidad de la asignación ya que puede interpretarse como el número esperado de péptidos con puntuación igual o superior a la del mejor péptido candidato.

Por último, *Mascot* es quizás el más popular de los motores de búsqueda a pesar de ser comercial y de que el algoritmo de correlación que usa nunca fue publicado. El programa calcula una puntuación expresada en términos probabilísticos llamada *ion score* que indica la probabilidad de que un número de coincidencias de picos hayan ocurrido aleatoriamente dado el número total de picos en el espectro y dada una distribución calculada de los valores *m/z* predichos para los candidatos

1.6.2. Otras estrategias de asignación péptido-espectro

Búsqueda basada en bibliotecas de espectros

Una alternativa posible a la búsqueda de espectros MS/MS usando espectros teóricos predichos computacionalmente a partir de bases de datos de secuencias consiste en buscar mediante comparación directa con otros espectros ya almacenados en una biblioteca de espectros. Estas bibliotecas se crean mediante la recopilación de espectros MS/MS observados e identificados en experimentos previos. Un nuevo espectro adquirido puede ser comparado directamente con los espectros de la biblioteca (que se encuentren dentro de un rango de tolerancia de masa permitida) y determinar así cual es la mejor coincidencia.

Al igual que en el caso de los motores de búsqueda basados en secuencia, existe un tipo específico de *software* que permite crear bibliotecas de espectros y realizar búsquedas usándolas como SpectraST ([Lam et al., 2007](#)), BiblioSpec () o X!Hunter, ()

Este tipo de aproximación supera a la búsqueda basada en secuencia en términos de velocidad, tasa de error y sensibilidad en la identificación de péptidos ([Lam2007](#)) Además, a los resultados obtenidos también se les puede aplicar los modelos de validación estadística desarrollados para las búsquedas basadas en

secuencia.

Sin embargo, en contrapartida, sólo es posible detectar aquellos péptidos que hayan sido previamente identificados y que se encuentren en la biblioteca de espectros

Identificación por secuenciación *de novo*

La secuenciación *de novo* (figura ??), a diferencia de las otras aproximaciones para interpretar espectros MS/MS, no requiere información adicional como las secuencias de las proteínas o espectros recopilados en experimentos previos. Por este motivo, la interpretación de espectros *de novo* es útil para detectar proteínas de organismos no secuenciados o procedentes de muestras de origen desconocido.

Existe también para este tipo de aproximación *software* que automatiza el proceso. Sin embargo su uso no se encuentra muy extendido ya que, para la gran cantidad de espectros obtenidos en un experimento típico de *shotgun*, el proceso es computacionalmente muy exhaustivo y requiere espectros MS/MS de gran calidad.

1.7. Evaluación estadística de las asignaciones PSM

Frecuentemente en un solo experimento de Proteómica *shotgun* se generan decenas de miles de espectros MS/MS. El procesamiento bioinformático automatizado de estos datos es por tanto un aspecto fundamental para la interpretación de los resultados. Por otra parte, no a todos los espectros MS/MS generados se les asigna un péptido, y a su vez, de todo el conjunto de PSM sólo una fracción son correctos, es decir el espectro corresponde realmente a la secuencia asignada. De hecho, en algunos experimentos realizados en instrumentos de baja resolución, los PSM incorrectos pueden llegar a suponer la mayoría ([Nesvizhskii, 2007](#)). Por eso, el desarrollo de métodos de evaluación de la calidad, en términos de confianza estadística, es una tarea crucial para filtrar los resultados generados.

1.7.1. Conceptos estadísticos básicos

El planteamiento general en el tratamiento estadístico en experimentos de Proteómica consiste en enfrentar dos hipótesis. La hipótesis nula (H_0) indica que el péptido (o proteína) está incorrectamente identificado. La hipótesis alternativa (H_1) indica lo contrario, que la asignación es correcta. Los tests estadísticos que se aplican enfrentan ambas hipótesis para aportar una medida de probabilidad estadística, generalmente la probabilidad de rechazar la hipótesis nula, es decir, de que la asignación sea correcta. (Figura 1.10). La población que se estudia puede ser la puntuación de todos los candidatos enfrentados a un espectro, o en términos globales, para todo un experimento, las puntuaciones de todos los PSM.

En la tabla de contingencia de la figura 1.10, U, V, T, y S corresponden a los Verdaderos Negativos, Falsos Positivos, Falsos Negativos y Verdaderos Positivos respectivamente. Con estos valores se pueden definir los conceptos de

- *sensibilidad* o Tasa de Verdaderos Positivos, TPR (*True Positive Rate*). Es la proporción de asignaciones consideradas correctas (por encima del umbral) entre el total de asignaciones correctas.

$$TPR = \frac{S}{T + S} \quad (1.5)$$

- *especificidad*. Es la proporción de asignaciones consideradas incorrectas entre el total de asignaciones incorrectas.

$$especificidad = \frac{U}{U + V} \quad (1.6)$$

- Tasa de Falsos Negativos, FNR (*False Negative Rate*). Es la proporción de asignaciones consideradas incorrectas entre el total de asignaciones correctas.

$$FNR = 1 - sensibilidad = \frac{T}{T + S} \quad (1.7)$$

- Tasa de Falsos Positivos, FPR (*False Positive Rate*). Es la proporción de asignaciones consideradas correctas entre el total de asignaciones incorrectas.

$$FPR = 1 - especificidad = \frac{V}{U + V} \quad (1.8)$$

1.7. Evaluación estadística de las asignaciones PSM

INTRODUCCIÓN

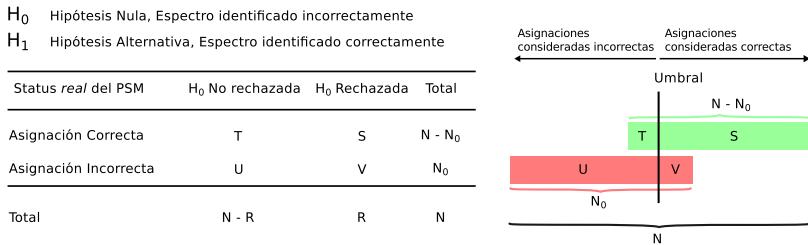


Figura 1.10: Tabla de contingencia. Contraste de hipótesis

- Tasa de Falsos Descubrimientos, FDR (*False Discovery Rate*). Es la proporción de asignaciones consideradas incorrectas entre el total de asignaciones consideradas (por encima del umbral).

$$FDR = \frac{V}{S + V} \quad (1.9)$$

1.7.2. Puntuaciones basadas en distribuciones de espectro individual y promedio

Distribución de espectro individual

Generalmente los motores de búsqueda aportan varios tipos de puntuación para cada PSM. Un tipo de puntuaciones se refiere a la calidad de cada asignación en particular, evalúa el grado de similitud entre el espectro empírico y el péptido asignado, el mejor de la lista de candidatos. (*Xcorr* en Sequest, *hyperscore* en X!Tandem o *ionScore* en Mascot). Pero además, a veces, aportan una puntuación indicativa de la calidad del PSM en relación a otros, al segundo mejor candidato (ΔCn de Sequest); o con respecto a una población del resto de candidatos que obtuvieron puntuaciones inferiores utilizando los parámetros estadísticos *e*-valor y *p*-valor.

Para ello, en primer lugar se selecciona el mejor péptido asignado a un espectro, es decir aquel candidato con la mejor puntuación, y a continuación se construye una distribución de las puntuaciones del resto de péptidos comparados con el espectro.

INTRODUCCIÓN

Esta distribución representa la hipótesis nula, la población de asignaciones PSM incorrectas.

El *p*-valor se calcula entonces relacionando la puntuación del mejor péptido con respecto a esta distribución (aleatoria) del resto de puntuaciones. Cuanto más alejada se sitúa la mejor puntuación del centro de la distribución mayor es la significatividad estadística del PSM. El *p*-valor, es por tanto, una medida de la probabilidad de que el mejor péptido candidato seleccionado sea asignado incorrectamente al espectro. Así, un *p*-valor bajo indicará una baja probabilidad de que el PSM haya sido asignado de forma incorrecta, es decir, es probablemente correcto.

El *e*-valor también se usa frecuentemente como medida de calidad en aproximaciones de espectro individual. Esá relacionado con el *p*-valor pero se interpreta como el número esperado de péptidos con puntuación igual o superior a la del mejor péptido candidato. X!Tandem calcula un *e*-valor obtenido empíricamente a partir de la distribución de espectro individual para cada PSM (Figura 1.11)

Ambos parámetros estadísticos, el *p*-valor y el *e*-valor, a diferencia del valor de puntuación original calculado por el motor de búsqueda, son independientes de la función de puntuación utilizada y por tanto suponen una medida más general de la calidad de cada PSM y son comparables en ensayos que usan distintos instrumentos, diferentes motores de búsqueda y parámetros (Nesvizhskii, 2010).

Algunos motores de búsqueda, además de su función de puntuación propia, como *hyperscore* en el caso de X!Tandem o *ion score* en el caso de Mascot también hacen uso de una distribución de espectro individual para calcular y proporcionar un *e*-valor para cada PSM.

Distribución promedio

En los experimento *shotgun* generalmente se obtienen miles de espectros MS/MS. Las medidas estadísticas de las distribuciones de espectro individual por tanto, no son suficientes. Incluso en el caso de que se requiera un *p*-valor muy bajo, (lo que implicaría una confianza estadística muy alta para un PSM en concreto) si se evalúan miles de espectros MS/MS podrían ocurrir PSM con *p*-valores igualmente bajos solo por azar. Por este motivo se utilizan estrategias de *corrección de test*

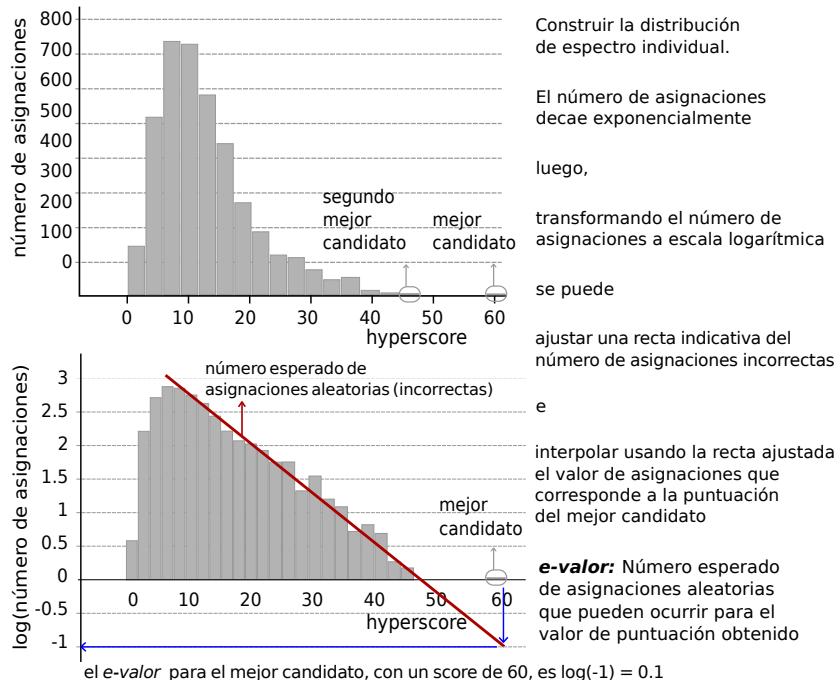


Figura 1.11: Estimación del e-valor

múltiple (*multiple test correction*) que re-ajustan los *p*-valores. Una aproximación muy utilizada, aunque produce resultados conservadores, es la corrección de Bonferroni (Abdi, 2007), que simplemente divide el *p*-valor por el número de veces que se repite el test. Así para un PSM con *p*-valor = 0,05 en un experimento en el que hay otros 10.000 PSM, el *p*-valor original habría de reajustarse a $0,05/10.000 = 5 \cdot 10^{-6}$.

Las distribuciones promedio, como muestra la Figura 1.12, son distribuciones de las mejores puntuaciones de todos los PSM de un experimento y permiten por tanto estimar otros parámetros estadísticos adicionales a nivel global, como la Tasa de Falsos Descubrimientos, FDR y la probabilidad de un PSM en particular en el

INTRODUCCIÓN

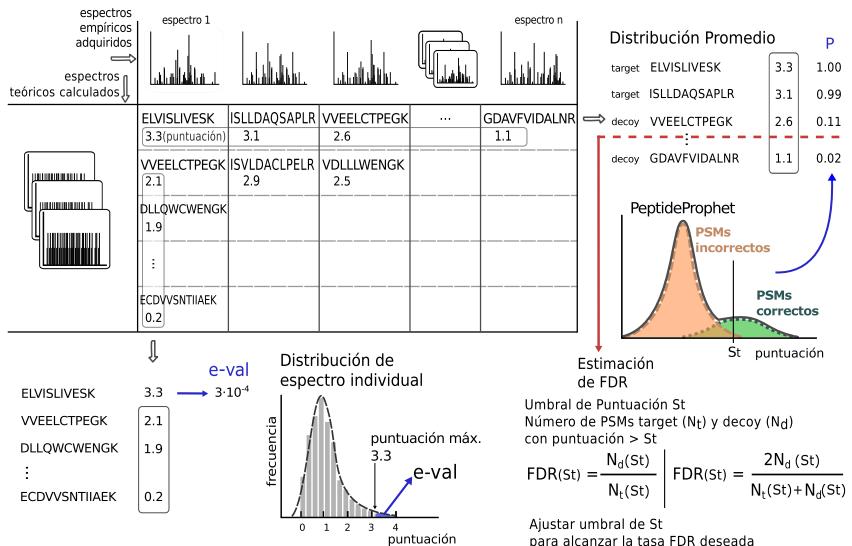


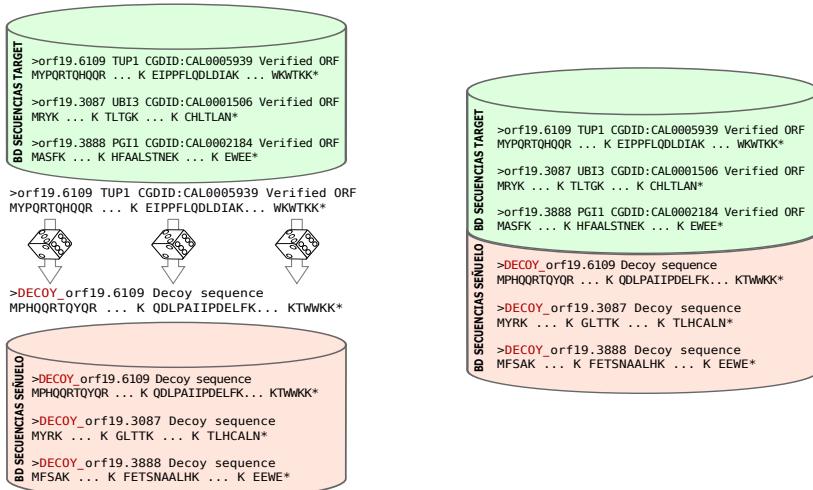
Figura 1.12: Distribuciones de Espectro Individual y Promedio

contexto global del experimento.

Es importante destacar que las aproximaciones que usan distribuciones de espectro individual son compatibles con las que usan distribuciones promedio, es decir, se puede realizar un análisis FDR global para un conjunto de PSM que han sido ordenados por *p*-valores o *e*-valores obtenidos individualmente.

1.7.3. Bases de datos señal y Tasa de Falsos Descubrimientos (FDR)

El tipo de evaluación estadística más ampliamente utilizada en experimentos de Proteómica *shotgun* es un tipo de corrección de test múltiples, la *Tasa de Falsos Descubrimientos (FDR, False Discovery Rate)* (Benjamini y Hochberg, 1995). Básicamente, el concepto de tasa FDR se refiere a la proporción de PSM incorrectos que se aceptan en todo el conjunto de PSM de un experimento para un umbral de puntuación (o de parámetro estadístico como el *p*-valor) fijado.

Figura 1.13: Construcción de una base de datos *señuelo*

Para la estimación de la tasa FDR (Figura 1.12), la estrategia utilizada consiste esencialmente en utilizar una base de datos llamada *señuelo* o *decoy* (Elias y Gygi, 2007). Es una aproximación sencilla pero efectiva que requiere que los espectros MS/MS sean comparados con espectros teóricos derivados de secuencias *señuelo*, que pueden ser generadas de varias formas pero que, en cualquier caso, son secuencias que no existen, no corresponden a ninguna proteína (Figura 1.13). La asignación de espectros MS/MS a estas secuencias *señuelo* permite recrear una hipótesis nula. Se puede tener la certeza de que los resultados de identificaciones correspondientes a secuencias *señuelo*, claramente etiquetadas en el fichero fasta, son identificaciones incorrectas. A continuación, para hacer las búsquedas, se puede añadir a la base de datos de secuencias reales (secuencias *target*) un número equivalente de secuencias *señuelo* y hacer que el motor de búsqueda use esta base de datos concatenada (*target-decoy*) del doble de tamaño que la original. O bien se pueden realizar dos búsquedas consecutivas, una utilizando la base de datos de secuencias *target* y otra a continuación utilizando la de secuencias *señuelo*.

Las secuencias *señuelo* pueden obtenerse mediante varios métodos (Elias y

INTRODUCCIÓN

(Gygi, 2007; Käll et al., 2008) La inversión de la secuencia de la proteína es un método sencillo que conserva la frecuencia media de cada aminoácido y permite generar siempre las mismas secuencias *señuelo* para sucesivas búsquedas. A cambio, el hecho de que no sea un orden aleatorio puede implicar que la población *señuelo* no refleje exactamente una hipótesis nula. También se pueden generar las secuencias de cada proteína de forma aleatoria. Esto también conserva las frecuencias de los aminoácidos, pero por otra parte, se elimina toda redundancia y se generarán por tanto un mayor número de péptidos *señuelo*. Otra opción es, en lugar de generar nuevas secuencias para cada proteína, crear péptidos *señuelo* de cada proteína dado el patrón de corte conocido de la enzima proteolítica utilizada. Esta opción tiene la ventaja de que los péptidos creados serán el mismo número y tendrán exactamente las mismas masas que las secuencias reales.

Una vez establecida esta hipótesis nula, la estrategia asume una idea básica central: la frecuencia con que los espectros MS/MS son asignados a secuencias *señuelo* sigue la misma distribución que la frecuencia con que los espectros son asignados incorrectamente a secuencias *target*.

Así, de forma general y dado que las bases de datos de secuencias *target* y *señuelo* tienen el mismo tamaño, el número de PSM incorrectos o Falsos Positivos (N_{inc} , aquellos espectros a los que se ha asignado incorrectamente una secuencia *target*) puede ser considerado equivalente al número de PSM *señuelo* (N_d , espectros a los que se ha asignado una secuencia *señuelo*). Con esto se puede estimar la tasa FDR como N_d/N_t , esto es, la proporción de PSM *señuelo*, N_d como sustituto conocido de N_{inc} , entre el total de secuencias *target* con puntuaciones superiores al umbral fijado, N_t . En ocasiones, cuando las búsquedas se hacen sobre la base de datos concatenada, para tener en cuenta que el tamaño es el doble que la original, la tasa FDR también puede calcularse como $2N_d/(N_t+N_d)$

Esta estimación general puede tener variantes. En el caso de que se realicen dos búsquedas independientes, una sobre la base de datos *target* y a continuación sobre la equivalente *señuelo*, la estimación de FDR como N_d/N_t resulta conservadora ya que N_d puede considerarse una sobre-estimación de N_{inc} . Esto se debe a que toda la población de espectros se compara con las secuencias *señuelo* a pesar de que algunos de los espectros podrían asignarse correctamente a una secuen-

cia *target*. Además, la mayoría de las funciones de puntuación tienden a otorgar puntuaciones más altas a PSM *señuelo* que a PSM *target* incorrectos por lo que la distribución de puntuaciones *señuelo* no es un reflejo preciso de la distribución de puntuaciones de los PSM incorrectos. Una forma de corregir este efecto consiste en estimar una aproximación previa de la fracción N_{inc} dentro de N_t considerando que la mayoría de los PSM con puntuaciones bajas son probablemente incorrectos (Käll et al., 2008) Así se puede incluir en la tasa FDR un factor de corrección definido por el porcentaje estimado de PSM *target* incorrectos (PIT): Si en N_t el 80% de los PSM son incorrectos, la tasa FDR calculada como N_d/N_t se multiplica por 0.8 para obtener un valor FDR más preciso (Por cada 100 PSM *señuelo* en el conjunto de PSM aceptado se estiman 80 PSM *target* incorrectos)

Las búsquedas utilizando una base de datos concatenada *target-decoy* son menos sensibles al efecto de sobre-estimación de N_d , sin embargo también producen un resultado FDR conservador. En este caso ya no se compara todo el conjunto de espectros con las secuencias *target* y *señuelo* por separado sino simultáneamente lo que produce un efecto de competición. Se puede considerar que las secuencias *target* y *señuelo* compiten por el espectro. Pero esto implica que a algunos espectros se les puede asignar una secuencia *señuelo* con una puntuación mayor a la que se obtiene al asignarles la secuencia *target* correcta. En tal caso se produce un aumento de N_d y una consiguiente reducción del número de PSM correcto y por tanto un incremento de FDR

Otra forma de mejorar la estimación de FDR es un algoritmo refinado (Navarro y Vázquez, 2009) que consiste en una búsqueda en bases de datos separadas teniendo en cuenta en conjunto las distribuciones de poblaciones de PSM *target* y PSM *decoy* y corrige el efecto de competición de las búsquedas en bases de datos concatenadas.

1.7.4. Modelos mixtos de probabilidad. Probabilidad Posterior

La estrategia de las bases de datos *señuelo* permite una estimación global de la tasa FDR pero no proporciona un valor de confianza estadística para cada PSM individual.

PeptideProphet (Keller et al., 2002) es un algoritmo de post-procesamiento (em-

INTRODUCCIÓN

pleado después de que el motor de búsqueda haya establecido una lista de PSM), el primero en implementar este tipo de análisis, que permite estimar la confianza en los péptidos identificados aportando un valor de probabilidad posterior.

Resumidamente, PeptideProphet recalcula las puntuaciones y emplea un método de Bayes empírico -un procedimiento de inferencia estadística en que la distribución *a priori* se estima a partir de los datos- para establecer un modelo mixto de probabilidad que integra las subpoblaciones de espectros correctamente asignados e incorrectamente asignados.

Primero, PeptideProphet recalcula las puntuaciones. Mediante análisis discriminante, las distintas puntuaciones aportadas por un motor de búsqueda son combinadas en un solo valor que maximiza la separación de asignaciones correctas e incorrectas. La puntuación discriminante S resulta de una función combinación ponderada de las puntuaciones x_1, x_2, \dots, x_s expresada de forma general:

$$S = F(x_1, x_2, \dots, x_s) = c_0 + \sum_{i=1}^s c_i x_i \quad (1.10)$$

donde la constante c_0 y el peso de las variables c_i son derivadas de forma que la proporción de la variación entre clases (asignaciones correctas e incorrectas) se maximiza con respecto a la variación dentro de cada clase.

Y como ejemplo específico para el caso de Sequest:

$$S = F_{SEQUEST}(Xcorr, \Delta C_n, SpRank) = c_0 + c_1 Xcorr + c_2 \Delta C_n + c_3 SpRank \quad (1.11)$$

Aunque en su versión original, el software requería una población de asignaciones correctas de referencia para estimar estas variables, posteriormente el algoritmo fue extendido para poder estimar los coeficientes de forma dinámica a partir de los datos en cada experimento(Ding et al., 2008; Ma et al., 2012).

PeptideProphet asume que la distribución de las puntuaciones recalculadas S puede explicarse como una combinación, una distribución mixta, en la que las asignaciones realmente correctas siguen una distribución $Normal(\mu, \sigma)$, y las asignaciones incorrectas, una distribución $Gamma(\alpha, \beta, \gamma)$. (Figura 1.14).

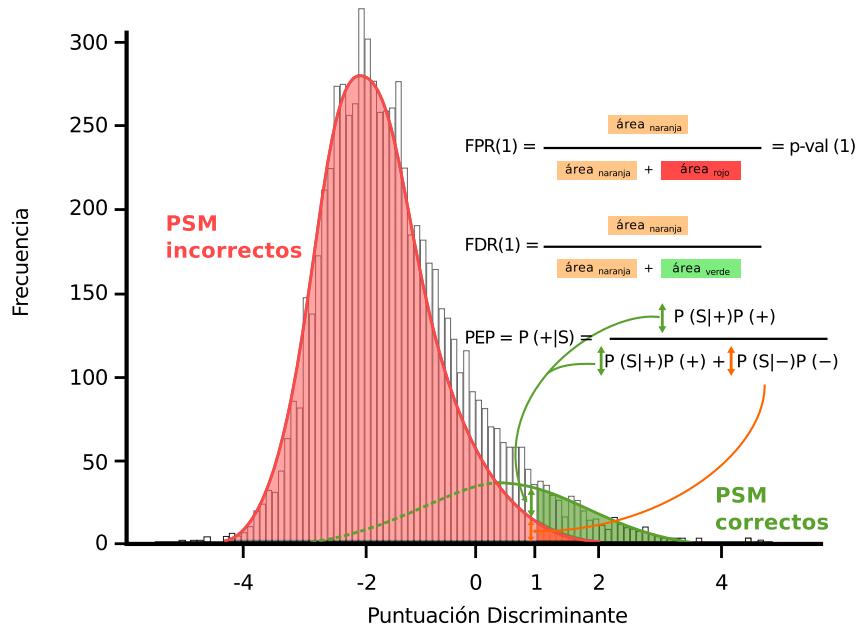


Figura 1.14: Estimación de PeptideProphet de las distribuciones de PSM incorrectos y correctos

Pero además de la puntuación discriminante hay otros parámetros que contribuyen a la separación de las poblaciones de PSM incorrectos y PSM correctos. Concretamente el número de extremos trípticos, NTT (*Number of Tryptic Termini*), el número de puntos de corte no efectuados, NMC (*Number of Missed Cleavages*) y el error en la masa del precursor, ΔM , tienen individualmente distribuciones diferentes para las asignaciones correctas e incorrectas (Choi y Nesvizhskii, 2008) lo que aporta una mejor definición de las distribuciones cuando son incorporados en un modelo mixto común.

A continuación, PeptideProphet usa un algoritmo de Esperanza-Maximización, EM (*Expectation-Maximization*) y el teorema de Bayes para estimar las distribuciones *Normal*, de PSM correctos, y *Gamma*, de PSM incorrectos; y calcular una probabilidad para cada asignación individual.

INTRODUCCIÓN

Para iniciar, el algoritmo requiere los parámetros π_0 (la proporción de asignaciones incorrectas en toda la población); μ, σ , parámetros que definen la *Normal*; y α, β, γ , que definen la *Gamma*.

En el primer paso -*E*- se usa el teorema de Bayes para estimar la probabilidad condicionada de cada puntuación de ser correcta:

$$P(+|S) = \frac{P(S|+)P(+)}{P(S|+)P(+) + P(S|-)P(-)} \quad (1.12)$$

donde $P(+|S)$ es la probabilidad de que el PSM con puntuación S sea correcta, $P(S|+)$ y $P(S|-)$ son las probabilidades condicionadas de una puntuación S entre las distribuciones correctas e incorrectas; y $P(+)$ y $P(-)$ son las probabilidades *a priori* de asignaciones correctas e incorrectas. (Figura 1.14). Esta probabilidad puede entenderse como la proporción de la la densidad de *Gamma* escalada por π_0 con respecto a la suma de las densidades *Gamma* y *Normal* escaladas π_0 y $1 - \pi_0$ para una puntuación S . O dicho con otras palabras, la probabilidad de que, teniendo una puntuación S , una asignación sea incorrecta con respecto a la probabilidad de tener esa puntuación. Este valor es la Probabilidad de Error Posterior, *PEP* que coincide con la FDR local.

Y en el siguiente paso -*M*-, una vez calculadas las probabilidades de cada PSM se recalculan los valores de los parámetros que describen las distribuciones.

Los pasos *E* y *M* se suceden iterativamente hasta la convergencia, el momento en que los parámetros estimados no difieren en valor absoluto de un error predefinido suministrado, ϵ .

Además de proporcionar probabilidades para cada PSM, PeptideProphet también calcula la FDR global. Para un umbral de puntuación t :

$$FDR(t) = \frac{P(-)P(S > t | -)}{P(S > t | -)P(-) + P(S > t | +)P(+)} \quad (1.13)$$

y otros parámetros como *p*-valor y la tasa de Falsos Positivos FPR. (Figura 1.14).

En un principio, cuando fue implementado (Keller et al., 2002), este modelo mixto para el cálculo de probabilidades no hacía uso de búsquedas realizadas usando la estrategia *target-señuelo* pues no se había generalizado aún. Cuando

comenzaron a emplearse este tipo de búsquedas PeptideProphet incorporó esta información haciendo que las puntuaciones correspondientes a péptidos *señuelo* sólamente puedan contribuir a la estimación de los parámetros que describen la distribución *Gamma* de PSM incorrectos. Esto permitió redefinir una versión semi-supervisada del algoritmo ([Choi y Nesvizhskii, 2008](#)) en el sentido de que la clase (PSM correctos e incorrectos) pasa a ser conocida para algunas pero no todas las asignaciones.

The Probability Ratio Method, Navarro et al

1.8. Inferencia de proteínas a partir de péptidos

En un experimento de Proteómica *shotgun*, desde el momento de la digestión de las proteínas, todo el análisis subsiguiente se realiza a nivel de péptidos. Esto, que permite que la estrategia *shotgun* pueda obtener un alto rendimiento en la identificación de péptidos, sin embargo provoca una dificultad adicional para ensamblar una lista de proteínas que presumiblemente se encuentran en la muestra analizada.

Uno de los principales motivos que complican la inferencia de las proteínas se refiere a la pérdida de la correspondencia péptido-proteína como consecuencia fundamentalmente de la detección de péptidos compartidos o *degenerados* cuyas secuencias están en diferentes proteínas. Esta dificultad, descrita como *el problema de la inferencia de proteínas* en ([Nesvizhskii y Aebersold, 2005](#)), a su vez se deriva de varios posibles escenarios. En algunos casos el procesamiento alternativo de intrones provoca la existencia de isoformas de una proteína en la muestra. De éstas, sólo algunas tienen en su secuencia péptidos trípticos exclusivos que permitan, en caso de ser detectados, concluir la presencia de la proteína fehacientemente. En otros casos, proteínas diferentes procedentes de una familia de genes (parálogos) poseen una alta homología de secuencia. Frecuentemente, a pesar de detectar un cierto número de péptidos, no se puede indicar la presencia de ninguna de las proteínas de la familia en particular.

En función de la presencia de estos péptidos compartidos y de péptidos exclusivos de cada proteína se han descrito métodos y nomenclaturas adecuadas para definir como *identificada* una proteína o lista de proteínas ([Nesvizhskii y Aebersold,](#)

INTRODUCCIÓN

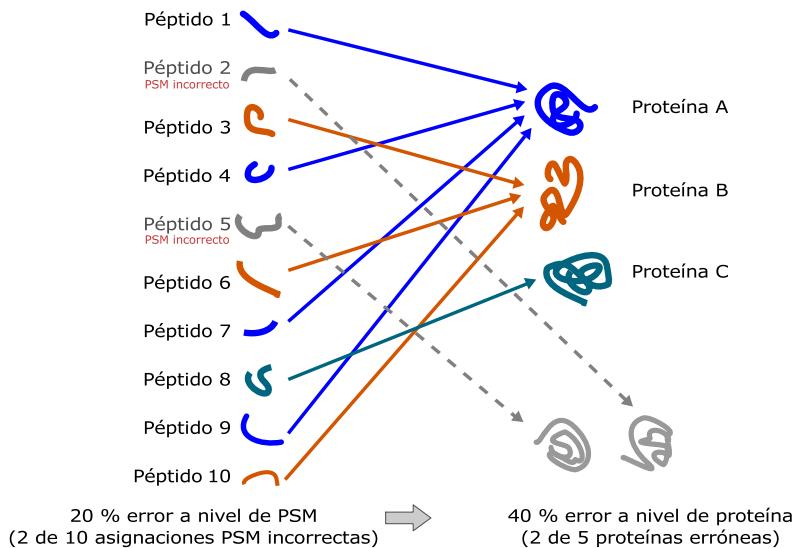


Figura 1.15: Agrupamiento no aleatorio de péptidos en proteínas

2005)

Un principio básico que ha sido muy utilizado es el de la llamada navaja de Occam ([Nesvizhskii et al., 2003](#)) consiste en presentar el menor número posible de proteínas que pueda explicar todos los péptidos observados y, en el caso de que varias proteínas estén representadas por varios péptidos compartidos (ninguno de ellos exclusivo de una proteína), presentar las proteínas en un grupo como una sola entrada de la lista.

Otra de las dificultades en la inferencia de proteínas es el fenómeno consistente en el agrupamiento dirigido de péptidos en sus correspondientes proteínas, lo que provoca una amplificación de las tasas de error. Como se observa en la figura 1.15, mientras que las identificaciones de péptidos correctas tienden a agruparse en un número pequeño de proteínas, los péptidos incorrectos, puesto que proceden de asignaciones aleatorias a entradas en la base de datos, suponen una proteína errónea por cada péptido.

Para combinar los péptidos en las proteínas originarias se pueden realizar va-

rias aproximaciones. La más sencilla consiste en seleccionar de todos sus PSM el de mejor puntuación o probabilidad y usarlo como puntuación de la proteína. Sobre este método básico se puede añadir además reglas como que al menos dos péptidos que superen un cierto umbral de puntuación deben contribuir a la inferencia de la proteína. A menudo se emplean variaciones de este método, fijando distintos umbrales y criterios, sin embargo una aproximación estadística puede ser más interesante, utilizando las probabilidades a nivel de PSM para combinarlas y obtener una probabilidad a nivel de proteína. La herramienta ProteinProphet ([Nevzihskii et al., 2003](#)) implementa este tipo de análisis. ProteinProphet refina las probabilidades de cada PSM previamente a combinarlas para tener en cuenta el problema del agrupamiento dirigido de péptidos en proteínas descrito, especialmente importante en proteínas representadas por un solo péptido. Este reajuste inicial consiste en tener en cuenta el número de péptidos hermanos, NSP (*Number of Sibling Peptides*) de forma que se penalizan las probabilidades de péptidos cuando sólo uno aporta evidencia para una proteína mientras que las probabilidades se reajustan aumentándose en casos en que muchos péptidos hermanos aporten evidencia a la presencia de la proteína. El reajuste teniendo en cuenta el parámetro NSP se realiza con una aproximación de Bayes empírica análoga a los modelos de PeptideProphet que estima las distribuciones de NSP entre las poblaciones de péptidos correcta e incorrecta:

$$NSP_i = \sum_{j \neq i} p_j$$

$$p'_i = \frac{p_i f_1(NSP_i)}{p_i f_1(NSP_i) + (1 - p_i) f_0(NSP_i)} \quad (1.14)$$

$$P(prot) = 1 - \prod_i (1 - p'_i)$$

donde $f_0(NSP)$ y $f_1(NSP)$ son las distribuciones NSP entre los péptidos incorrectos y correctos.

1.9. Herramientas adicionales de post-procesamiento y validación a nivel de péptido y proteína

Algunas herramientas bioinformáticas de post-procesamiento de resultados de identificaciones como PeptideProphet y ProteinProphet han demostrado ser de gran utilidad proporcionando un medio de calcular probabilidades y tasas de error de forma muy precisa. Sin embargo, la modelización que emplean a veces resulta demasiado excesiva, especialmente en casos de listas de espectros e identificaciones muy grandes (Reiter et al., 2009).

Por otra parte, frecuentemente interesa desde el punto de vista experimental, alcanzar una cobertura lo más amplia posible del proteoma objeto de estudio utilizando para ello una aproximación combinada con varios motores de búsqueda.

La herramienta de *software* iProphet (Shteynberg et al., 2011) fue desarrollada para resolver este tipo de necesidades, implementando una modelización más completa de todas las fuentes de información en un experimento de Proteómica *shotgun*. Si PeptideProphet aporta probabilidades a nivel de PSM, esta extensión refina las probabilidades aportándolas a nivel de secuencia peptídica única, es decir combina todas las probabilidades de los PSM que se refieren a la misma secuencia. Para ello, el programa reajusta (mejorando o penalizando) los valores de salida de PeptideProphet(de forma análoga a como ProteinProphet lo hace con el parámetro NSP) usando cinco fuentes adicionales de información: El número de búsquedas, NSS (*Number of Sibling Searches*), que aumenta las probabilidades de una secuencia peptídica si es identificada por múltiples motores de búsqueda; el número de espectros repetidos, NRS (*Number of Replicate Spectra*), que tiene en cuenta si hay muchos espectros que son asignados a la misma secuencia con alta probabilidad; el número de réplicas, NSE (*Number of Sibling Experiments*), que aumenta las probabilidades de péptidos que son repetidamente identificados a partir de espectros obtenidos en distintos experimentos (asumiendo que son réplicas o protocolos similares); el número de iones de un péptido, NSI (*Number of Sibling Ions*), que aumenta la probabilidad de un péptido si es encontrado en distintos estados de carga; y por último el número de instancias modificadas, NSM (*Number of Sibling Modifications*), que actúa de forma similar a NSI, aumenta la probabilidad si

se encuentra una instancia modificada y sin modificar de un péptido.

Por otra parte, las herramientas hasta ahora descritas para el control de la tasa de errores, ya sea mediante la estrategia de búsqueda en bases de datos señuelo o bien con el control estadístico que aportan los modelos mixtos de probabilidad (*PeptideProphet*), permiten controlar la tasa FDR a nivel de PSM. Sin embargo obtener los valores de FDR a nivel de proteína implica un nivel adicional de complejidad. Dado que una proteína se considera identificada cuando contiene un conjunto de PSM que a su vez pueden ser correctos o no, el error, como muestra la figura 1.15, se propaga de forma especialmente acusada en experimentos que generan grandes conjuntos de datos (decenas de miles de espectros) (Reiter et al., 2009). Por eso la estimación de FDR a nivel de proteína ha de tener en cuenta que los PSM falsos positivos y los PSM verdaderos positivos tienen distribuciones diferentes. Mientras que los primeros apuntarán aleatoriamente a entradas de toda la base de datos, los segundos solo corresponden al subconjunto de proteínas presentes en la muestra. Esto hace que en la práctica las tasas de error para proteínas sean mayores que para PSM.

1.10. Proteómica dirigida. SRM/MRM

La Proteómica *shotgun*, cuyo objetivo es detectar la mayor cantidad posible de proteínas en una muestra, se denomina en ocasiones por ello, Proteómica *de descubrimiento*. En esto, esencialmente, la *Proteómica dirigida* se distingue de las técnicas de *shotgun*, en el objetivo. Esta metodología no pretende identificar una gran cantidad de proteínas diferentes en la muestra, sino que intenta identificar y, opcionalmente también cuantificar, una proteína o un grupo de proteínas de interés seleccionadas *a priori*. De ahí el nombre *Proteómica dirigida*.

La técnica que se utiliza para llevar a cabo experimentos de Proteómica dirigida se denomina *SRM*, *Selected Reaction Monitoring* o también, frecuentemente utilizado como sinónimo, *MRM*, *Multiple Reaction Monitoring*.

La Proteómica dirigida, basada en técnicas SRM, está actualmente emergiendo y popularizándose como un complemento ideal de las técnicas de *shotgun*.

Básicamente, SRM proporciona unas propiedades muy interesantes en experi-

INTRODUCCIÓN

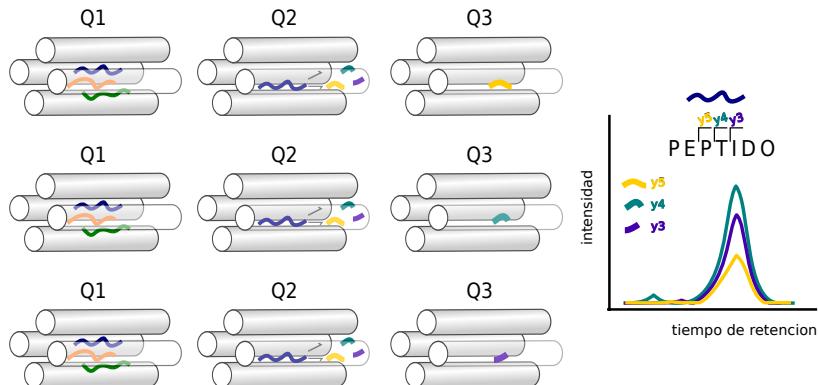


Figura 1.16: Adquisición y reconstrucción de la señal en un experimento SRM

mentos en que se requiere que un grupo de proteínas, por ejemplo biomarcadores o proteínas constituyentes de una red o ruta particular, sean detectadas y cuantificadas de una forma precisa y reproducible en diferentes muestras que se quiere comparar.

Originalmente desarrollada para detectar y cuantificar pequeñas moléculas como metabolitos o drogas (Referencia), las primeras aplicaciones de SRM al campo de la Proteómica comenzaron en 2003 ([Gerber et al., 2003](#)), 2004 ([Kuhn et al., 2004](#)).

El principio fundamental en el que se basa la técnica SRM consiste en aprovechar la capacidad de espectrómetros de masas de tipo triple cuadrupolo para actuar como filtros de masas de analitos a dos niveles consecutivos, el de un péptido precursor, y el de los fragmentos que se generan tras ser éste fragmentado. Este doble filtro es, idealmente, muy representativo del péptido, y por extensión, de la proteína originaria.

La señal que se genera en el instrumento al medir la cantidad de estos fragmentos, junto con la información del tiempo de retención cromatográfica permite reconstruir un pe

1.11. Repositorios públicos de Proteómica a gran escala y dirigida

1.11.1. Repositorios públicos de Proteómica a gran escala y dirigida

Los resultados generados en experimentos de Proteómica han de ser convenientemente mostrados y compartidos con la comunidad científica

1.11.1.1. PRIDE

La base de datos PRIDE (*Protein Identifications Database*), creada en el Instituto Bioinformático Europeo en Inglaterra (?) es el repositorio más extenso de datos de espectrometría de masas. Almacena los resultados originales enviados por los investigadores usando para ello su propio formato PRIDE XML. Además se han desarrollado herramientas que facilitan la creación y el envío de los resultados en este formato.

1.11.1.2. PeptideAtlas

El proyecto PeptideAtlas surgió en 2005 en el Instituto de Biología de Sistemas, Seattle, Washington. (Desiere et al., 2006). A diferencia de PRIDE, donde los resultados enviados no son reanalizados de ninguna forma sino que se mantienen como los usuarios los enviaron, PeptideAtlas sí cuenta con un flujo de validación de los resultados. Este análisis, denominado TPP (*Trans Proteomics Pipeline*) (Deutsch et al., 2010) emplea secuencialmente los programas descritos PeptideProphet, ProteinProphet y iProphet principalmente, para asegurar la calidad y robustez de los datos de identificaciones mostrados. Inicialmente contaba con datos de proteínas humanas y posteriormente un gran número de especies se han incorporado. El trabajo titulado *A Candida albicans PeptideAtlas* (Vialas et al., 2013) forma parte del proyecto desde 2012.

1.12. Formatos de archivos usados en espectrometría de masas y Proteómica

En el proceso de análisis de datos que sigue a la adquisición experimental de espectros se requiere un uso intensivo de *software*, desde la asignación de secuencias peptídicas a los espectros hasta la elaboración de listas de proteínas identifi-

INTRODUCCIÓN

cadas y evaluación estadística de los resultados. Existe una gran variedad de este tipo de programas, que sirven de apoyo a cada uno de estos pasos en el proceso de análisis

En términos muy generales se puede distinguir *software* abierto, que la comunidad bioinformática ha desarrollado en respuesta a las necesidades de compartir, inspeccionar y generar ficheros sin las restricciones que imponen las licencias; y *software* privativo desarrollado *ad hoc* por las compañías fabricantes de espectrómetros de masas para sus instrumentos. En este sentido la iniciativa HUPO-PSI ha adquirido un papel muy importante en la elaboración y difusión de formatos abiertos que puedan servir de estándar para toda la comunidad

En una clasificación más precisa, el software puede clasificarse en función de la etapa del análisis al que sirven de ayuda.

- *Formatos que recogen la salida de los espectrómetros de masas*

Este es un tipo de formatos muy diverso. Depende básicamente, de la forma cuando la frecuencia en que se escanea cada fragmento es superior a la resolución del instrumento la señal se registra como picos con una forma y anchura precisas. Este tipo de adquisición es el *modo continuo o perfil*. Los instrumentos registran los espectros en modo continuo de forma predefinida, pero frecuentemente son sometidos a un procesamiento por un algoritmo que extrae los picos detectados como parejas de valores *m/z* e intensidad. Esto se denomina adquisición de *datos centroide*

Entre los formatos desarrollados por los fabricantes podemos encontrar aquellos para los que toda la información de los espectros se encuentra en un solo archivo, aquellos para los que la información se divide en un par de archivos y aquellos con múltiples archivos para cada adquisición de espectros. Así, para los instrumentos Thermo Scientific el formato es del primer tipo, toda la información es codificada en archivos con extensión .RAW (datos perfil o centroide a elección). Los instrumentos AB-Sciex en su mayoría (excepto los TOF-TOF) pueden generar archivos del segundo tipo, en pares donde un archivo con extensión .wiff contiene los metadatos y un archivo .wiff.scan contiene los espectros. Por último, para algunos instrumentos de Waters y

1.12. Formatos de archivos usados en espectrometría de masas y Proteómica

INTRODUCCIÓN

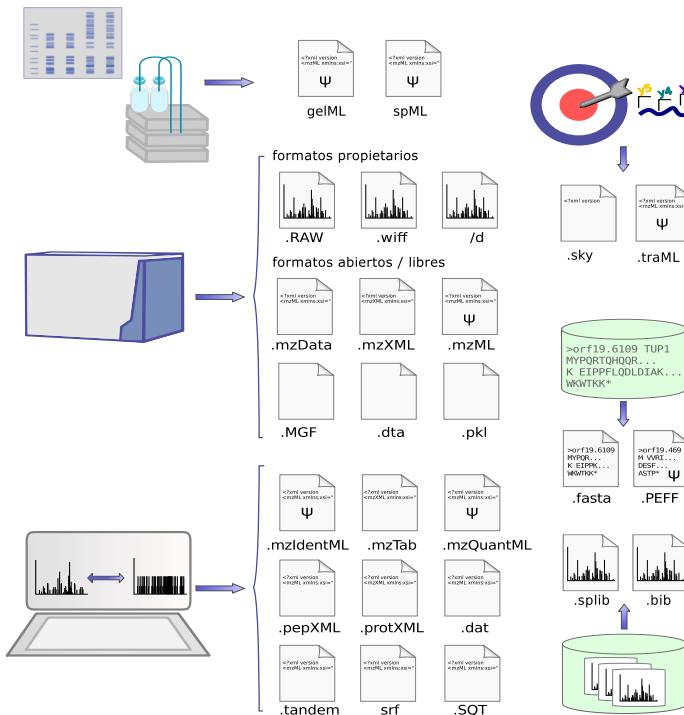


Figura 1.17: Visión general de formatos comúnmente usados en cada etapa de un experimento de Proteómica

Agilent, se obtienen múltiples archivos agrupados en carpetas con extensión *.d* o *.raw*.

Sin embargo, el hecho de que estos formatos sean codificados en binario junto con la disponibilidad de las librerías de lectura proporcionadas por los fabricantes restringida únicamente al sistema operativo MS Windows frenó el desarrollo de nuevas herramientas de lectura y manipulación de este tipo de archivos. En ese contexto, aparecieron los primeros formatos de texto (XML) para codificar toda la información de salida de MS, mzXML (?) y mzData () que posteriormente fueron unificados en el estándar HUPO-PSI mzML (Deu-

INTRODUCCIÓN

tsch, 2010).

Por último, una solución intermedia, ideada previamente a la aparición de los formatos basados en XML descritos, es la creación de ficheros de texto simples con una simple lista de los picos obtenidos para cada ión precursor y sus fragmentos. Los formatos de archivo con extensión *.pk1* *.dta* o *.ms2* contienen espectros independientes en cada fichero, los *.MGF* pueden contener múltiples espectros en un solo fichero.

- *Formatos que recogen el resultado de las búsquedas*

Este tipo de formatos se encuentra generalmente muy ligado al *software* empleado para generar los resultados, es decir el motor de búsqueda. Sequest, comenzó usando los formatos *.out* y *.SQT* pero posteriormente ha desarrollado los *.SRF* y *.MSF*. X!Tandem y OMSSA generan archivos basados en XML *.tandem* y *.omx* respectivamente.

Para independizar el motor de búsqueda usado del formato obtenido se creó el formato *.pepXML* que además permite el análisis por las herramientas de TPP. *pepXML*, cuya unidad de información básica es el PSM, es el formato que lee y escribe PeptideProphet, para ProteinProphet, se creó *.protXML*, que contiene la lista de proteínas y sus péptidos asignados.

Sin embargo, de nuevo *pepXML* y *protXML*, aunque muy populares, también estaban ligados al flujo de análisis de TPP. Y de nuevo, HUPO-PSI ideó un nuevo formato estándar para recoger toda información derivada del resultado de búsquedas y análisis independientemente de su origen, el mzIdentML ([Jones et al., 2012](#)). Además, HUPO-PSI también ha desarrollado mzTab, una versión alternativa simplificada que no se basa en XML sino en texto separado por tabulador.

- *Formatos que almacenan bibliotecas de espectros*

Los motores de búsqueda que usan bibliotecas de espectros, como SpectraST (parte de TPP) ([Lam et al., 2007](#)) requieren generalmente un formato que contenga espectros consenso, una combinación de los espectros y el péptido que se les ha asignado, así como otras anotaciones y metadatos. El

Instituto Nacional para Estándares y Tecnología americano, NIST (*National Institute for Standards and Technology*) distribuye bibliotecas de espectros en formato *.msp*. SpectraST produce el formato *.splib*; y X!Hunter y BiblioSpec *ASL* y *.blib* respectivamente.

- *Formatos que almacenan secuencias*

Los motores de búsqueda basados en secuencia requieren que se les suministren secuencias de cada proteína (y también, para cada una de ellas su correspondiente secuencia señalero). Para ello el tipo de formato más usado es el celeberrimo FASTA. Sin embargo no es el único, HUPO-PSI ha creado el formato PEFF, que mejora a FASTA añadiendo reglas sobre cómo ha de expresarse la cabecera de cada secuencia. Esta sintaxis definida facilita la tarea al *software* que lee los ficheros.

- *Formatos específicos para Proteómica dirigida*

En Proteómica dirigida, podemos encontrar dos tipos de formatos. Entre los que se usan como fuente de entrada de información, para indicar al espectrómetro las listas de transiciones, es decir, que precursores y fragmentos ha de monitorizar, los más empleados son *.sky*, empleado por el programa Skyline; y el estándar HUPO-PSI *TraML* (Deutsch et al., 2012). Para los resultados de análisis por SRM, el programa Skyline usa su propio formato, *.skyd*, basado en XML y otros programas como mProphet (Reiter et al., 2011) usan sus propios formatos de texto separado por tabulador.

1.13. *Candida albicans* como organismo modelo

Candida albicans es un hongo patógeno oportunista que se encuentra comúnmente como residente comensal, inocuo, en las mucosas gastrointestinal y urogenital en un alto porcentaje de la población. Sin embargo, su cualidad de patógeno oportunista implica que, en ocasiones, propiciadas generalmente por un sistema inmune debilitado en el hospedador, puede proliferar y diseminarse provocando infecciones, candidiasis, de gravedad variable, desde afecciones mucocutáneas leves hasta infecciones sistémicas severas que pueden incluso llegar a ser letales.

INTRODUCCIÓN

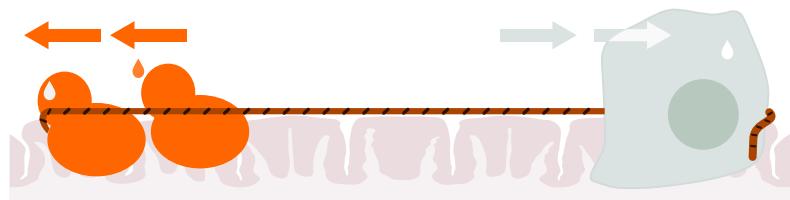


Figura 1.18: Status de equilibrio entre las células de *Candida albicans* y células del sistema inmune en las mucosas

Los principales factores de virulencia con los que cuenta *Candida albicans* para proliferar y diseminarse causando infecciones son su capacidad de cambiar su morfología de una forma levaduriforme a una forma hifal, siendo la primera más adecuada para la diseminación y la segunda para penetrar e invadir tejidos; y su capacidad de adhesión gracias a la acción de proteínas adhesinas (Calderone y Fonzi, 2001)

Desde el punto de vista de la Proteómica, se han realizado muy diversos estudios usando Candida Estudios basados en 2D-PAGE para investigar las diferencias en el proteoma de la forma de levadura y la forma de hifa

YUJUUU: COMPLETAR ESTO VALEEE!!?

OBJETIVOS

Objetivos

- Desarrollo de una aplicación web respaldada por una base de datos para recoger, almacenar y mostrar resultados de identificación de proteínas procedentes de experimentos de proteómica relacionados con *Candida albicans*
- Recopilación de resultados de espectrometría de masas para la creación de un PeptideAtlas, o Atlas Peptídico, de *Candida albicans*
- Creación de una base de datos para recoger y almacenar métodos de proteómica dirigida (MRM) empleados para detectar proteínas en el contexto del consorcio español dedicado al mapeo cromosoma-centrífico en el proyecto proteoma humano (HPP)

DESARROLLO DE UNA APLICACIÓN WEB PARA
RESULTADOS DE IDENTIFICACIONES DE
PROTEÓMICA DE *Candida albicans*

*My primary goal when designing Ruby was to
have fun programming.*

Yukihiro Matsumoto

Proteopathogen, a protein database for studying Candida albicans - host interaction

Vital Vialás, Rubén Nogales-Cadenas, César Nombela, Alberto Pascual-Montano,
Concha Gil

Proteomics 2009, 9, 4664-4668

TECHNICAL BRIEF

Proteopathogen, a protein database for studying *Candida albicans* – host interaction

Vital Vialás¹, Rubén Nogales-Cadenas², César Nombela¹, Alberto Pascual-Montano² and Concha Gil^{1,3}

¹ Departamento de Microbiología II, Facultad de Farmacia, Universidad Complutense de Madrid, Madrid, Spain

² Departamento de Arquitectura de Computadores y Automática, Facultad de Ciencias Físicas, Universidad Complutense de Madrid, Madrid, Spain

³ Unidad de Proteómica UCM-Parque Científico de Madrid, Facultad de Farmacia, Universidad Complutense de Madrid, Madrid, Spain

There exist, at present, public web repositories for management and storage of proteomic data and also fungi-specific databases. None of them, however, is focused to the specific research area of fungal pathogens and their interactions with the host, and contains proteomics experimental data. In this context, we present Proteopathogen, a database intended to compile proteomics experimental data and to facilitate storage and access to a range of data which spans proteomics workflows from description of the experimental approaches leading to sample preparation to MS settings and peptides supporting protein identification. Proteopathogen is currently focused on *Candida albicans* and its interaction with macrophages; however, data from experiments concerning different pathogenic fungi species and other mammalian cells may also be found suitable for inclusion into the database. Proteopathogen is publicly available at <http://proteopathogen.dacya.ucm.es>

Received: January 13, 2009

Revised: June 25, 2009

Accepted: July 2, 2009

Keywords:

Candida albicans / Database / Host / MS / Microbiology / Pathogen

Candida albicans is an opportunistic pathogenic fungus, which can be found as a component of the usual flora in human mucoses. Although it does not normally cause disease in immunocompetent colonized hosts, in the case of immunosuppressed patients *Candida* cells can over-proliferate and become pathogenic. Cells in yeast form (oval cells) may produce hyphae, penetrate tissues and eventually cause invasive candidiasis. At present, the frequency of this fatal opportunistic mycosis continues to be distressing and, unfortunately, solution is hindered by the reduced effectiveness and serious side effects of the few available drugs,

the appearance of antifungal-drug resistance, and the lack of accurate and prompt diagnostic procedures [1].

Addressing proteomic studies involving the way *Candida* interacts with immune cells is thus essential in order to improve our comprehension of the process of infection and represents the primary step of investigation that could lead to future development of diagnosis methods, vaccines and antifungal drugs [2–5].

Experimental techniques in proteomics have quickly evolved in such a way that nowadays we have to deal with vast amounts of complex data originated by the combination of multi-dimensional separation techniques and MS analysis together with the bioinformatics software reports [6]. Existing public repositories for management and storage of proteomic data such as World 2-D PAGE [7], the Proteome Database System for Microbial Research 2-D PAGE [8], or PRIDE [9]; and fungi-specific databases such as BioBase MycoPathPD [10], Candida Genome Database (CGD) [11] or Candida DB [12] are very popular and useful tools. However, none of them deals with proteomic experimental

Correspondence: Dr. Concha Gil, Departamento de Microbiología II, Facultad de Farmacia, Universidad Complutense, Plaza de Ramón y Cajal s/n, 28040 Madrid, Spain

E-mail: conchagil@farm.ucm.es

Fax: +34-913941745

Abbreviations: CGD, Candida Genome Database; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; PDB, Protein Data Bank

data related to the specific research area of fungal pathogens and their interaction with the host. In this context, we present Proteopathogen, a protein database, currently focused on the *C. albicans* – macrophage interaction model – which enables a framework for the access and submission of proteomic workflow data, from description of the experimental approaches leading to sample preparation to MS settings and identification-supporting peptides. Through its interface web site, the database can easily be queried to allow an efficient browsing through all the stored data, improving the quality of eventual analysis of MS results.

Regarding the compilation of information used to populate the database, data from three different studies were considered suitable to be present in Proteopathogen. The first two correspond to published works relating to proteomics of the *Candida* – macrophage interaction [2, 3], where the former reports 66 different *C. albicans* identified proteins and the latter, 38 murine macrophage proteins. The third study represents an analysis of the *C. albicans* plasma membrane proteome [13]. It compiles a set of experiments aimed at extraction and identification of membrane proteins and a set of experiments intended to obtain enrichment in glycosylphosphatidylinositol-anchored surface proteins, which have been reported to be involved in cell wall biogenesis, cell-cell adhesion and interaction with the host [14].

In all cases, protein identifications lists are collected together with the pertinent experimental context specified by descriptions of the experimental approaches, MS settings and peptides supporting identification for each of the proteins (Table 1).

Along with the experimental information, and in order to provide a deeper view of the data, complementary information is retrieved from public web repositories. In the case of *C. albicans* proteins, identifiers, synonyms, aminoacid sequence of the translated open reading frame, *Saccharomyces cerevisiae* orthologs, *Gene Ontology* (GO) annotation, pathway annotations and scientific literature references were obtained from CGD [11], whereas in the case of murine macrophage proteins, the equivalent information was obtained from UniProt KnowledgeBase [15] and the Mouse Genome Database [16]. Additionally, pathways annotations were retrieved from Kyoto Encyclopedia of Genes and Genomes (KEGG)

Pathway Database [17] and structure information from the Protein Data Bank (PDB) [18].

Concerning the architecture of the software, the back-end layer consists of a MySQL database managed by the web application development framework Ruby on Rails that sets up structure and relations of data, handles queries to the database and displays the user web-based interface.

The experimental context is addressed in Proteopathogen in a hierarchical manner, where a main general approach, which may correspond to a published article, is characterized by a description or title, authors, target species and Pubmed identifier when available; and experiments within it, are in turn, characterized by the description of the particular experiment, the date when it was performed and number of identified proteins.

Information on one particular protein is split into several sections in Proteopathogen. *Protein Basic Information* displays the UniProt accession number, description, species, evidence for the existence, standard gene name, organism-specific database identifiers, yeast orthologs for *Candida* proteins and human orthologs for mouse proteins and sequence. The Section 2 lists experiments in which the particular protein has been identified. Where available, one or more of the following sections will be displayed as well: the table entitled GO showing GO annotations along with the pertinent scientific references, the KEGG Pathways and CGD Pathways tables rendering annotations from KEGG and CGD respectively, and PDB, a table specifying structural information. Where no PDB identifiers are found for *C. albicans* proteins, *S. cerevisiae* orthologs are used instead, and similarly, when a PDB identifier cannot be found for mouse proteins, the human ortholog is used.

In all cases, proteins are unambiguously related to their corresponding experiment, thus enabling a relation to the data concerning experimental parameters of identification and identification-supporting peptides. This data comprise, on the one hand, common MS settings for all proteins identified in the particular experiment, including search database, MS type, analysis software, digestion enzyme, fixed aminoacid modifications, variable modifications and maximum allowed number of miscleavages; and on the other hand, particular parameters and peptides list for each protein, including number of matched peptides, score,

Table 1. Overview of the stored data in Proteopathogen as well as their published evidences

References	Description of experimental approach	Species	#Protein identifications
[2]	<i>C. albicans</i> differentially expressed proteins after 3 h interaction with RAW 264.7 murine macrophages. 2-D silver-stained gel. MS/MS (MALDI/TOF-TOF)	<i>C. albicans</i>	66
[3]	Proteins identified from cytoplasmic extracts of RAW 264.7 cells after 45 min interaction with <i>C. albicans</i>	<i>Mus. musculus</i>	38
[13]	Identification of Glycosyl phosphatidil inositol (GPI)-anchored membrane proteins Identification of membrane proteins	<i>C. albicans</i>	292 1273

observed peptide mass, calculated peptide mass, start and end coordinates, number of missed cleavages and the sequence of the peptide.

The web interface to Proteopathogen offers multiple ways to query the database. Through the *Browse Experiments* search option, a list containing all sets of experimental approaches is displayed. In its turn, one particular experiment can be browsed through all the proteins identified in it.

The *Search* form may be used in different manners. Queries for one particular protein can be performed by supplying one of the multiple supported identifiers, namely standard gene names, Candida feature name, Candida DB identifiers, CGD identifiers, MGI identifiers and UniProt accession numbers. Free text queries can be performed as

well, which will retrieve a list of proteins showing coincidences in the description field of the Proteopathogen protein entry. As an additional feature, peptide sequences can also be searched for retrieving in this case, proteins in any experiment having the searched sequence in any of the identification-supporting peptides. Wild characters ("*") and Boolean operators are supported for free text queries and for peptide sequence queries.

In order to enhance interactivity and collaboration with users, a submission form is included in the web interface to allow the upload of more proteomic experimental approaches as long as they concern the topics addressed in Proteopathogen. Sequential steps request from the user the following information: a description of the experimental context, a related protein list, MS parameters

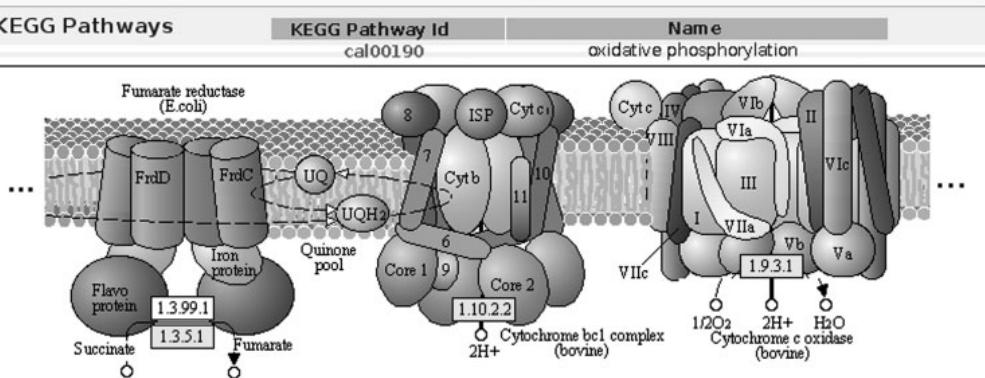
Basic Information	<p>UniProt Accession: Q59QS9 Description: Protein described as ubiquinol-cytochrome-c reductase ... Species: <i>Candida albicans</i> Existence: Verified Standard gene name: QCR2 CGD ID: CAL0003458 CDB ID: CA2055 S. cerevisiae ortholog: QCR2 Synonyms: orf19.10167, IPF24692.1, IPF6978.2... Sequence</p>	<p>Gene Ontology Annotations</p> <table border="1"> <thead> <tr> <th>Ontology</th> <th>Id</th> <th>Description</th> <th>Evidence</th> <th>Ref</th> </tr> </thead> <tbody> <tr> <td>BP</td> <td>GO:0006122</td> <td>mitochondrial, electron transport</td> <td>IEA</td> <td>CGD paper</td> </tr> <tr> <td>BP</td> <td>GO:0009060</td> <td>aerobic respiration</td> <td>IEA</td> <td>CGD paper</td> </tr> <tr> <td>MF</td> <td>GO:0008121</td> <td>ubiquinol-cyt-c reductase activity</td> <td>IEA</td> <td>PubMed</td> </tr> <tr> <td>CC</td> <td>GO:0005624</td> <td>membrane fraction</td> <td>IDA</td> <td>PubMed</td> </tr> <tr> <td>CC</td> <td>GO:0005750</td> <td>mitochondrial respiratory chain</td> <td>IEA</td> <td>CGD paper</td> </tr> </tbody> </table>	Ontology	Id	Description	Evidence	Ref	BP	GO:0006122	mitochondrial, electron transport	IEA	CGD paper	BP	GO:0009060	aerobic respiration	IEA	CGD paper	MF	GO:0008121	ubiquinol-cyt-c reductase activity	IEA	PubMed	CC	GO:0005624	membrane fraction	IDA	PubMed	CC	GO:0005750	mitochondrial respiratory chain	IEA	CGD paper
Ontology	Id	Description	Evidence	Ref																												
BP	GO:0006122	mitochondrial, electron transport	IEA	CGD paper																												
BP	GO:0009060	aerobic respiration	IEA	CGD paper																												
MF	GO:0008121	ubiquinol-cyt-c reductase activity	IEA	PubMed																												
CC	GO:0005624	membrane fraction	IDA	PubMed																												
CC	GO:0005750	mitochondrial respiratory chain	IEA	CGD paper																												
Experiments	<p>MS Experiment: Method D. Dounce homogeniser protoplast breaking and 12–60% sucrose gradient. LC-LTQ</p>	<p>PDB</p> <table border="1"> <thead> <tr> <th>PDB id</th> <th>Summary</th> </tr> </thead> <tbody> <tr> <td>1kb9b</td> <td>Yeast cytochrome bc1 complex</td> </tr> </tbody> </table> 	PDB id	Summary	1kb9b	Yeast cytochrome bc1 complex																										
PDB id	Summary																															
1kb9b	Yeast cytochrome bc1 complex																															
Protein Identification parameters for QCR2, Q59QS9	<table border="1"> <thead> <tr> <th>P pro</th> <th>Score</th> <th>Coverage</th> <th>Mw</th> <th>Peptide Hits</th> </tr> </thead> <tbody> <tr> <td>8.34 E-39</td> <td>40.21</td> <td>16.04</td> <td>39557.2</td> <td>8 (8 0 0 0 0)</td> </tr> </tbody> </table>	P pro	Score	Coverage	Mw	Peptide Hits	8.34 E-39	40.21	16.04	39557.2	8 (8 0 0 0 0)	KEGG Pathways																				
P pro	Score	Coverage	Mw	Peptide Hits																												
8.34 E-39	40.21	16.04	39557.2	8 (8 0 0 0 0)																												
Peptides List supporting identification for QCR2, Q59QS9	<table border="1"> <thead> <tr> <th>MH +</th> <th>Delta m</th> <th>Delta cn</th> <th>Z</th> <th>P pep</th> <th>Xc</th> <th>Sp</th> <th>Peptide</th> </tr> </thead> <tbody> <tr> <td>1115.64185</td> <td>-0.29329</td> <td>0.47</td> <td>2</td> <td>7.25E-08</td> <td>2.30</td> <td>807.5</td> <td>K.LSVIINNAGSK.T</td> </tr> <tr> <td>1550.79077</td> <td>-0.82267</td> <td>0.71</td> <td>2</td> <td>2.10E-08</td> <td>4.17</td> <td>1863.4</td> <td>K.SVAESVSSSALSEAVKA</td> </tr> </tbody> </table>	MH +	Delta m	Delta cn	Z	P pep	Xc	Sp	Peptide	1115.64185	-0.29329	0.47	2	7.25E-08	2.30	807.5	K.LSVIINNAGSK.T	1550.79077	-0.82267	0.71	2	2.10E-08	4.17	1863.4	K.SVAESVSSSALSEAVKA	<p>KEGG Pathway Id cal00190</p> <p>Name oxidative phosphorylation</p> 						
MH +	Delta m	Delta cn	Z	P pep	Xc	Sp	Peptide																									
1115.64185	-0.29329	0.47	2	7.25E-08	2.30	807.5	K.LSVIINNAGSK.T																									
1550.79077	-0.82267	0.71	2	2.10E-08	4.17	1863.4	K.SVAESVSSSALSEAVKA																									

Figure 1. Use case: Search for *C. albicans* ubiquinol-cytochrome-c reductase QCR2. The different sections in the result comprise information on protein description and identifiers, experiments in which it has been identified, GO annotation, KEGG and CGD pathway annotation and structural information from PDB.

and identification-supporting peptides lists. These data are subject to revision prior to eventual insertion into Proteopathogen by the database curators. Besides, the whole relational database and the MS data reports are available for download at the web site.

All the information that is retrievable from Proteopathogen when queried for one particular protein is shown in Fig. 1 for the specific case of ubiquinol-cytochrome-*c* reductase QCR2 of *C. albicans* which has been reported to show antigenic properties in human [19].

The *Protein Basic Information* section displays the Uniprot accession number, a brief description of the protein as stated at CGD, evidence for its existence, standard gene name, feature name, CGD and Candida Database identifiers, yeast ortholog gene name, synonyms and sequence.

The Section 2 lists all the experiments in which QCR2 has been identified. All of them belong to the same general approach aimed at purification of membrane proteins. In every case, the corresponding links to the MS identification parameters and supporting peptides are displayed as well. This experimental data are shown in Fig. 1 for identification of QCR2 in the experiment described as “Method D. Dounce homogenizer protoplast breaking and 12–60% sucrose gradient. LC-LTQ”.

The section entitled *GO annotations* shows terms related to the electron transport chain, but more interestingly, it also shows an *inferred from direct assay* (IDA) annotation to the term *membrane fraction* [20], which fits to the fact that the protein is identified in five of the methods aimed at purification of membrane proteins.

KEGG Pathways table provides a link to the KEGG Pathway entry for *Oxidative phosphorylation*, and provides the feature to show in place the image corresponding to the map from KEGG. CGD Pathways displays an analogous link to the pathway entry at CGD that, in this case, is named *aerobic respiration (cyanide sensitive)–electron donors*.

Finally, in the PDB section, there are four structure images available along with links to the PDB entries, corresponding to a cytochrome bc1 complex from *S. cerevisiae*. Orthologs were used since no structure could be found for the *Candida* protein.

In conclusion, Proteopathogen represents, up to date, the first public web-based repository for proteomics data related to studies involving *C. albicans* pathogenicity and its interaction with immune system cells in the host. Moreover, it enables a framework for public access and submission of this type of data and it is intended to be more actively populated in the near future, including data from different pathogenic fungi and mammalian cells, becoming a reference database in its field. Unlike other protein identification databases, Proteopathogen is focused to a specific topic but, at the same time, includes a wide range of data including descriptions of the experimental contexts, information on proteins such as GO and pathway annotations, structural information and detailed MS parameters. Therefore, Proteopathogen will contribute to save time and facilitate

analysis of proteomic workflow reports for researchers interested in this area.

The authors are grateful to César Vicente from the Computer Architecture Department, Complutense University of Madrid for his excellent technical assistance. This work was supported by BIO 01989-2006 from the Comisión Interministerial de Ciencia y Tecnología (CYCIT, Spain), DEREPLICOBIANA – CM from Comunidad Autónoma de Madrid, and REIPI, Spanish Network for the Research in Infectious Diseases, RD06/0008/1027 from the Instituto de Salud Carlos III. The Proteomics work was carried out in the Proteomics Unit UCM-Parque Científico, a member of the National Institute for Proteomics PROTEORED, funded by Genoma España. APM and RNC are partially supported by Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006 and PS-010000-2008-1.

The authors have declared no conflict of interest.

References

- [1] Calderone, R. A. (Ed.), *Candida and Candidiasis*, ASM Press, Washington D.C 2002.
- [2] Fernández-Arenas, E., Cabezón, V., Bermejo, C., Arroyo, J., et al., Integrated genomic and proteomic strategies bring new insight into *Candida albicans* response upon macrophage interaction. *Mol. Cell. Proteomics* 2007, 6, 460–478.
- [3] Martínez-Solano, L., Nombela, C., Molero, G., Gil, C., Differential protein expression of murine macrophages upon interaction with *Candida albicans*. *Proteomics* 2006, 6, 133–144.
- [4] Pitarch, A., Nombela, C., Gil, C., *Candida albicans* biology and pathogenesis: insights from proteomics. *Methods Biochem. Anal.* 2006a, 49, 285–330.
- [5] Pitarch, A., Nombela, C., Gil, C., Contributions of proteomics to diagnosis, treatment, and prevention of candidiasis. *Methods Biochem. Anal.* 2006b, 49, 331–361.
- [6] Monteoliva, L., Albar, J. P., Differential proteomics: an overview of gel and non-gel based approaches. *Brief Funct. Genomic Proteomics* 2004, 3, 220–239.
- [7] Hoogland, C., Mostaguir, K., Appel, R. D., Lisacek, F., The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server. *J. Proteomics* 2008, 71, 245–248.
- [8] Pleissner, K. P., Eifert, T., Buettner, S., Schmidt, F. et al., Web-accessible proteome databases for microbial research. *Proteomics* 2004, 4, 1305–1313.
- [9] Martens, L., Hermjakob, H., Jones, P., Adamski, M. et al., PRIDE: the proteomics identifications database. *Proteomics* 2005, 5, 3537–3545.
- [10] Csank, C., Costanzo, M. C., Hirschman, J., Hodges, P. et al., Three yeast proteome databases: YPD, PombePD, and CalPD (MycopathPD). *Methods Enzymol.* 2002, 350, 347–373.

- [11] Arnaud, M. B., Costanzo, M. C., Skrzypek, M. S., Binkley, G. *et al.*, The Candida Genome Database (CGD), a community resource for *Candida albicans* gene and protein information. *Nucleic Acids Res.* 2005, 33, 358–363.
- [12] Rossignol, T., Lechat, P., Cuomo, C., Zeng, Q. *et al.*, CandidaDB: a multi-genome database for Candida species and related Saccharomycotina. *Nucleic Acids Res.* 2008, 36, 557–561.
- [13] Cabezón, V., Llama-Palacios, A., Nombela, C., Monteoliva, L., Gil, C., Analysis of *Candida albicans* plasma membrane proteome. *Proteomics* 2009, 9, in press, DOI: 10.1002/pmic.200800988.
- [14] Plaine, A., Walker, L., Da Costa, G., Mora-Montes, M. *et al.*, Functional analysis of *Candida albicans* GPI-anchored proteins: roles in cell wall integrity and caspofungin sensitivity. *Fungal Genet. Biol.* 2008, 45, 1404–1414.
- [15] UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2008, 36, 190–195.
- [16] Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., *et al.*, The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* 2008, 36, 724–728.
- [17] Kanehisa, M., Araki, M., Goto, S., Hattori, M. *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 2008, 36, 480–484.
- [18] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G. *et al.*, The Protein Data Bank. *Nucleic Acids Res.* 2000, 28, 235–242.
- [19] Pitarch, A., Abian, J., Carrascal, M., Sanchez, M. *et al.*, Proteomics-based identification of novel *Candida albicans* antigens for diagnosis of systemic candidiasis in patients with underlying hematological malignancies. *Proteomics* 2004, 4, 550–559.
- [20] Insenser, M., Nombela, C., Molero, G., Gil, C., Proteomic analysis of detergent-resistant membranes from *Candida albicans*. *Proteomics* 2006, 6, S74–S81.

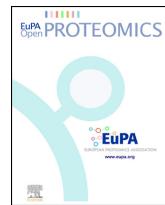
Proteopathogen2: A database and web tool to store and display proteomics identification results in the mzIdentML standard

Vital Vialas, Concha Gil

EuPA Open Proteomics 2015, IN PRESS



ELSEVIER

Available online at www.sciencedirect.com**ScienceDirect**journal homepage: <http://www.elsevier.com/locate/euprot>

Proteopathogen2, a database and web tool to store and display proteomics identification results in the mzIdentML standard[☆]

Vital Vialas ^{a,b,*}, Concha Gil ^{a,b}

^a Departamento de Microbiología II, Universidad Complutense Madrid (UCM), Spain

^b Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain

ARTICLE INFO

Article history:

Available online xxx

We want to dedicate this work to Juan Pablo's memory, for his dedication to standardization in proteomics and inspiring work in this field.

Keywords:

mzIdentML

Candida albicans

Web application

Database

Proteopathogen

ABSTRACT

The Proteopathogen database was the first proteomics online resource focused on experiments related to *Candida albicans* and other fungal pathogens and their interaction with the host. Since then, the HUPO-PSI standards were implemented and settled, and the first large scale *C. albicans* proteomics resource appeared as a *C. albicans* PeptideAtlas. This has enabled the remodeling of Proteopathogen to take advantage and benefit from the use of the HUPO-PSI adopted format for peptide and protein identification mzIdentML and continue offering a centralized resource for *C. albicans*, other fungal pathogens and different cell lines proteomics data.

© 2015 The Authors. Published by Elsevier B.V. on behalf of European Proteomics Association (EuPA). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The opportunist pathogenic fungus *Candida albicans*, under usual circumstances, is a harmless resident commensal in human mucous membranes of a large percentage of the population. However, taking advantage of weakened host immune defenses, for instance in immunocompromised cancer or AIDS patients, it may switch to its pathogenic status, overproliferating and becoming thus the main etiological agent of candidiasis, one of the most prevalent and costly types of fungal infections in global terms.

Proteomics studies have been addressed to study this commensal to pathogenic transition by approaching the dimorphic, yeast form to hyphal form switch [1,2], by specifically aiming at the study of some other clinically relevant biological processes such as apoptosis [3–5] or biofilm formation [6]; or targeting sets of proteins that interact first with the host like surface exposed and secreted proteins [6,7].

However, until recently, the resulting proteomics identification datasets were sparse and disseminated. The Proteopathogen database [8] was the first public online proteomics data repository specifically focused on experiments aimed at the study of *C. albicans* and other fungal species

[☆] This new Proteopathogen database and web tool is public online at <http://proteopathogen2.dacya.ucm.es>.

* Corresponding author at: Departamento de Microbiología II, Universidad Complutense Madrid (UCM), Spain. Tel.: +34 913941743.
E-mail address: vivialas@ucm.es (V. Vialas).

<http://dx.doi.org/10.1016/j.euprot.2015.04.002>

2212-9685/© 2015 The Authors. Published by Elsevier B.V. on behalf of European Proteomics Association (EuPA). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

pathogenic traits. Since no standard format for peptide and protein identification results was available, Proteopathogen was developed to compile and display identification lists in different tabulated text formats depending on the software used to generate and process the results.

At that time, the HUPO – Proteomics Standards Initiative (PSI) already had a trajectory striving to highlight the importance of standardization and providing formats that would comply with MIAPE (*Minimum Information About a Proteomics Experiment*) guidelines as reviewed in Ref. [9]. Some *de facto* standard formats existed like mzXML and pepXML [10], but the advent, years later, of the HUPO-PSI approved formats for mass spectrometry output data [11] and for identification results [12] among others, surfaced the efforts and claims by the community to finally adopt formats to facilitate data comparison, exchange and verification. This also inspired and boosted the development of an assortment of format conversion tools and libraries [13,14], and stand-alone software for visualization of the content of the files in standard formats [15] but, most importantly for the purpose of this work, enabled the possibility for Proteopathogen to benefit from the mzIdentML adopted standard for identification results, incorporating it as the input data format and using it as inspiration for information display.

More recently, the most comprehensive, up to the current date, online *C. albicans* proteomics data repository was developed and integrated in PeptideAtlas [16]. These publicly available *C. albicans* results have been used to establish a new version of Proteopathogen with a solid foundation.

In this background, we present here a revisited Proteopathogen database and web based tool adapted to read and display peptide and protein identification data based upon the mzIdentML format. It is the first online database specifically developed to map and store the contents of files in mzIdentML, it has been initially populated with the *C. albicans* PeptideAtlas identification results and it is publicly accessible at <http://proteopathogen2.dacya.ucm.es/>.

2. Materials and methods

The original identification result files were obtained from PeptideAtlas repository datasets PAe001976, PAe001977, PAe001978, PAe001979, PAe001980, PAe001981, PAe001982, PAe001983, PAe001984, PAe001985, PAe001986, PAe001987, PAe001988, PAe001989, PAe002110, and PAe002111.

As described in Ref. [16] the data sets come from a range of experiments including yeast to hypha transition assays, membrane protein extractions and a set of phosphoprotein enrichment approaches. In all cases, cells from the clinical isolates SC5314 were grown in YPD medium. For obtaining cells in hyphal form, either heat-inactivated fetal bovine serum or Lee medium pH 6.7 was used. As for the mass spectrometry, spectra were acquired in different set ups and platforms in a data-dependent manner. A summary of the experiments set ups and conditions is shown in Table 1.

Consistently with the PeptideAtlas project principles, the MS output files were processed through the Trans Proteomic Pipeline. The steps involved, first, sequence database searching using X! Tandem with k-score [18] and a custom sequence database obtained from *Candida* Genome Database [19] with

Table 1 – Summary of experiments, MS output files, instrument and PeptideAtlas datasets.

Type of dataset	Number of MS output files	Instrument	Peptide Altas datasets
<i>Candida albicans</i> culture with SILAC labeling, digested protein extracts enriched in phosphopeptides IMAC/TiO2	57	Orbitrap XL, Orbitrap Velos	PAe001976 PAe001977 PAe001978 PAe001979 PAe001980 PAe001984 PAe001985 PAe001986 PAe001987 PAe001988 PAe001989 PAe001983
<i>Candida albicans</i> total protein extract, 2 Triple-TOF runs, 2 µg and 4 µg HYPHAL form and yeast form total protein extracts	2	Triple-TOF	
	8	Orbitrap Velos	PAe002110 PAe002111
LTQ membrane proteins [17]	3	LTQ	PAe001981
LTQ proteins from acidic subproteome [1]	8	LTQ	PAe001982

appended decoy counterparts and common contaminants for peptide-to-spectrum matching and FDR assessment. Then the post-processing validation tools PeptideProphet [20], ProteinProphet [21] and iProphet [22] provided filtered lists of peptides and proteins with high probabilities. And finally FDR was computed for different probability thresholds.

Each of the PeptideAtlas repository datasets consists on the MS output spectra files and a set of pepXML and protXML files with lists of high confidence peptide and proteins respectively. These were combined, independently for each dataset, by means of a custom script written in the Ruby scripting language (available in supplemental data) to create mzIdentML files (mzIdentML version 1.1.0) with the merged information. In order to check the files were generated correctly and ensure data quality they were all validated (semantic and MIAPE-compliant validation) with mzidValidator [15].

A completely new MySQL relational database was implemented *ad hoc* to map elements in the mzIdentML files as depicted in Fig. 1 (schema available in supplemental data). Then, using the Ruby scripting language (version 2.0.0) and the Rails web application development framework (version 4.0.0) a script was created to parse the data in the mzIdentML files, store the relevant elements in the corresponding tables (available in supplemental data) and eventually create the web application to display the data.

3. Results and discussion

A total number of sixteen mzIdentML files, corresponding to each of the PeptideAtlas repository datasets, grouped into five different experiments were compiled and used to initially populate the Proteopathogen database. These account

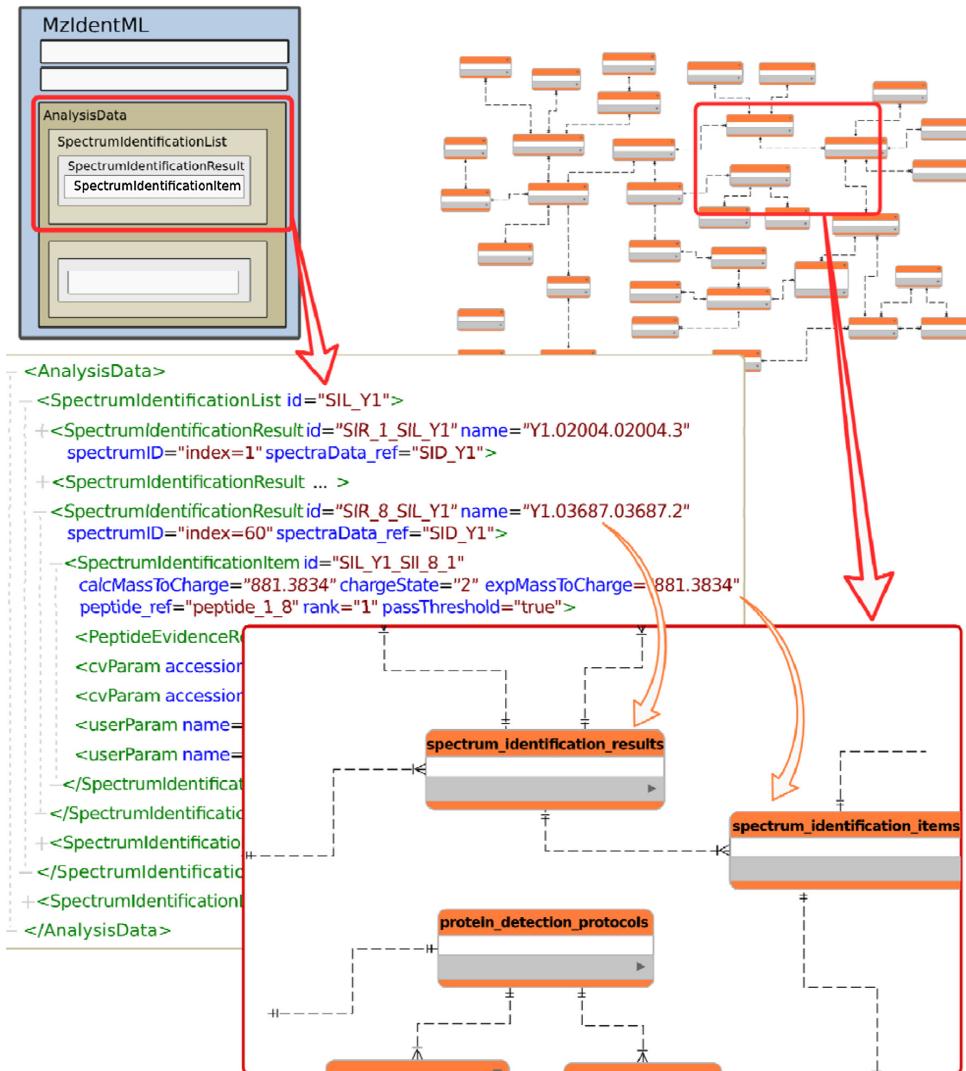


Fig. 1 – mzIdentML to database mapping. The MySQL schema was specifically designed to accommodate elements from the mzIdentML format. Figure shows the one-to-many relationship between the `<SpectrumIdentificationResult>` and `<SpectrumIdentificationItem>` elements.

for approximately 22,000 distinct peptides and 2600 different proteins that can be queried and viewed through the web interface.

Precisely, a stringent FDR cut-off at the PSM level set at 0.005, yields 21,883 peptides with 0.0024 FDR (peptide level) and 2577 proteins with 0.0170 FDR (protein level) as computed with Mayu, a software specifically designed to estimate accurate protein level error rates in large datasets [23] (see supplemental Table 1).

The mzIdentML contents can be browsed for each file in Proteopathogen in a means inspired by the structure in the format, particularly that under the `<AnalysisData>` element

containing the datasets generated by the analyses. That is, for each mzIdentML file, shown in its experimental context, a user can select either the spectrum identification information (corresponding to the `<SpectrumIdentification>` element) and view its related information, the search protocol, search database and the list of every peptide to spectrum assignment; or the protein detection (corresponding to the `<ProteinDetection>` element) showing the list of peptides grouped into the inferred original proteins (Fig. 2).

Notably, the information Proteopathogen displays will depend on how complete the original mzIdentML files are. For instance, for files including the optional `<Fragmentation>`

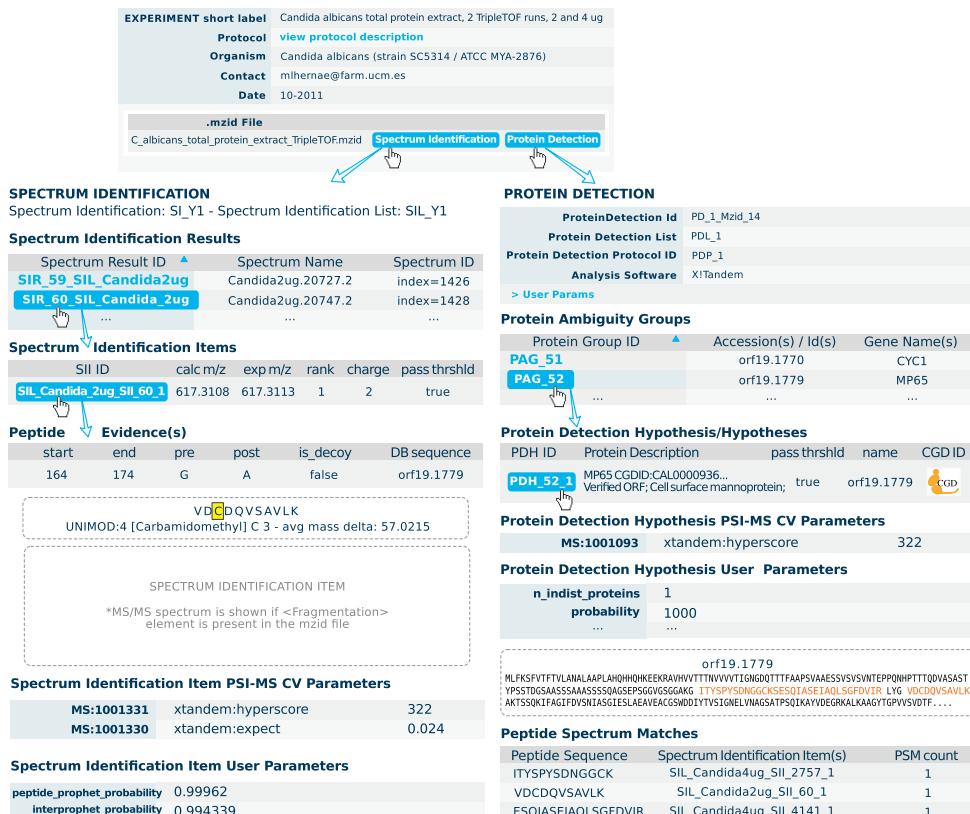


Fig. 2 – Information displayed in the web interface. Proteopathogen displays two main sets of information for the selected mzIdentML file. The spectrum identification section shows how for each spectrum there is a list of possible identification results, each having its peptide evidence, i.e. a sequence at a particular position in a protein sequence. The particular selected peptide is shown in its protein context in the protein detection section, which displays the complete list of the inferred proteins for the selected mzIdentML file with links to Candida Genome Database (CGD).

element under <SpectrumIdentificationItem>, Proteopathogen will display an annotated and interactive MS/MS spectrum. In addition, the optional <cvParam> and <userParam> elements, that describe and annotate with controlled vocabularies and user-defined information respectively different elements throughout the file, might be more or less profuse depending on the software that created them.

In addition to browsing through the contents of the stored mzIdentML files, Proteopathogen implements a query system yielding global results. That is, for a specific queried protein name, as found in the <ProteinDetectionHypothesis> name attribute, the search results display all the distinct peptide sequences found mapped into the protein sequence, regardless of the experiment in which they were identified, and the supporting spectra for each peptide sequence, while keeping track of the original <SpectrumIdentification> and mzIdentML file. A peptide sequence may also be searched, obtaining, when found, the corresponding protein, or group of proteins, and the set of supporting spectra, again in global scope.

The use of the Ruby scripting language, unlike other compiled languages (Java, C/C++) that are commonly used in other software used to visualize proteomics file formats, enables a quick, easy to implement, flexible manner of parsing complex XML files, and creation and manipulation of objects that have to be stored in a very precise order in a database. In addition, the argument of speed in computationally intensive tasks in favor of compiled languages is getting blurry nowadays with the array of XML parsing libraries that are continuously developed and improved for scripting languages. The type of solution implemented in Proteopathogen is a DOM (Document Object Model) parser, that creates an in-memory tree representation of the whole XML hierarchy. Arguably, a parser of the type SAX (Simple API for XML parsing) would perform better in terms of speed for large files but as trade-off, leaping back and forth in search of cross-referenced elements, as is the case in mzIdentML, would be difficult or even impossible to implement. Nevertheless, future work in the direction of a SAX implementation of the parser and a comparison in

performance with respect to the current one, would be of great interest.

Proteopathogen will greatly benefit from the adoption of mzIdentML as input data format. Any proteomics experiment on *C. albicans*, or any fungal pathogen–host interaction, as long as they are provided in valid (semantically valid and MIAPE-compliant) mzIdentML (version 1.1.0), will be welcome to be integrated in the database. To that purpose, users provided with login credentials may submit their files either through a simple upload form in the web application or transfer them using a specifically set up FTP server. Finally, the Rails framework for web application development will take care of any scalability issues with ease and allow for any kind of visualization improvements.

4. Conclusions

The Proteopathogen web application and database has been completely rebuilt to accommodate and display *C. albicans*, or any fungal pathogen for that matter, proteomics identification results in the HUPO-PSI adopted format for peptide and protein identification mzIdentML. This makes it the first public online database specifically designed to store the information contained in these types of files and display its contents following an analogous structure.

Transparency document

The Transparency document associated with this article can be found in the online version.

Acknowledgements

This work has been financially supported by project BIO2012-31767 Ministerio de Economía y Competitividad, Spain, PROPMT (S2010/BMD-2414) from the Comunidad Autónoma de Madrid, REIPI, Spanish Network for the Research in Infectious Diseases (RD12/0015/0004), and PRB2 (PT13/0001/0004) from the ISCIII. VV held a research contract associated to project BIO2012-31767. The authors are grateful to Daniel Tabas-Madrid and Alberto Pascual Montano from CNB-CSIC for outstanding support and assistance in setting up the web application.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.euprot.2015.04.002.

REFERENCES

- [1] Montelola L, Martínez-López R. Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition. *J Proteome Res* 2010;10:502–17, <http://dx.doi.org/10.1021/pr100710g>.
- [2] Gow N, van de Veerdonk F. *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization. *Nat Rev* 2011;10:112–22, <http://dx.doi.org/10.1038/nrmicro2711>.
- [3] Madeo F, Herker E, Wissing S. Apoptosis in yeast. *Curr Opin Microbiol* 2004;7:655–60, <http://dx.doi.org/10.1016/j.mib.2004.10.012>.
- [4] Fernández-Arenas E, Cabezón V. Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction. *Mol Cell Proteomics* 2007;6:460–78, <http://dx.doi.org/10.1074/mcp.M600210-MCP200>.
- [5] Ramsdale M. Programmed cell death in pathogenic fungi. *Biochim Biophys Acta* 2008;1783:1369–80, <http://dx.doi.org/10.1016/j.bbamcr.2008.01.021>.
- [6] Vialás V, Perumal P, Gutierrez D, Ximénez-Embún P, Nombela C, Gil C, et al. Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells. *Proteomics* 2012, <http://dx.doi.org/10.1002/pmic.201100588>.
- [7] Gil-Bona A, Llama-Palacios A, Parra CM, Vivanco F, Nombela C, Montelola I, et al. Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in *Candida albicans*. *J Proteome Res* 2014, <http://dx.doi.org/10.1021/pr5007944>.
- [8] Vialás V, Nogales-Cadenas R, Nombela C, Pascual-Montano A, Gil C. Proteopathogen, a protein database for studying *Candida albicans*–host interaction. *Proteomics* 2009;9:4664–8, <http://dx.doi.org/10.1002/pmic.200900023>.
- [9] Martínez-Bartolomé S, Binz P-A, Albar JP. The Minimal Information about a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative. *Methods Mol Biol* 2014;1072:765–80, http://dx.doi.org/10.1007/978-1-62703-631-3_53.
- [10] Deutscher E. File formats commonly used in mass spectrometry proteomics. *Mol Cell Proteomics* 2012;11:1612–21, <http://dx.doi.org/10.1074/mcp.R112.019695>.
- [11] Martens L, Chambers M, Sturm M. mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011;10, <http://dx.doi.org/10.1074/mcp.R110.000133>, R110.000133.
- [12] Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* 2012;11, <http://dx.doi.org/10.1074/mcp.M111.014381>, M111.014381.
- [13] Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 2012;30:918–20, <http://dx.doi.org/10.1038/nbt.2377>.
- [14] Griss J, Reisinger F, Hermjakob H, Vizcaíno JA. jmsReader: a Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics* 2012;12:795–8, <http://dx.doi.org/10.1002/pmic.201100578>.
- [15] Ghali F, Krishna R, Lukasse P, Martínez-Bartolomé S, Reisinger F, Hermjakob H, et al. Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML. *Mol Cell Proteomics* 2013;12:3026–35, <http://dx.doi.org/10.1074/mcp.O113.029777>.
- [16] Vialás V, Sun Z, Loureiro Y, Penha CV, Carrascal M, Abián J, et al. A *Candida albicans* PeptideAtlas. *J Proteomics* 2013, <http://dx.doi.org/10.1016/j.jprot.2013.06.020>.
- [17] Cabezón V, Llama-Palacios A, Nombela C, Montelola I, Gil C. Analysis of *Candida albicans* plasma membrane proteome. *Proteomics* 2009;9:4770–86, <http://dx.doi.org/10.1002/pmic.200800988>.
- [18] MacLean B, Eng J, Beavis R, McIntosh M. General framework for developing and evaluating database scoring algorithms

- using the TANDEM search engine. *Bioinformatics* 2006;22:2830–2, <http://dx.doi.org/10.1093/bioinformatics/btl379>.
- [19] Costanzo MC, Arnaud MB, Skrzypek MS, Binkley G, Lane C, Miyasato SR, et al. The Candida Genome Database: facilitating research on *Candida albicans* molecular biology. *FEMS Yeast Res* 2006;6:671–84, <http://dx.doi.org/10.1111/j.1567-1364.2006.00074.x>.
- [20] Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J Proteome Res* 2008;7:254–65, <http://dx.doi.org/10.1021/pr070542g>.
- [21] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;75:4646–58.
- [22] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* 2011;10, M111.007690–M111.007690, <http://dx.doi.org/10.1074/mcp.M111.007690>.
- [23] Reiter L, Claassen M, Schrimpf SSP, Jovanovic M, Schmidt A, Buhmann JM, et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* 2009;8:2405–17, <http://dx.doi.org/10.1074/mcp.M900317-MCP200>.

CREACIÓN DE UN PEPTIDEATLAS DE *Candida albicans*

*What evidence should satisfy you? Evidence
that is publicly recorded and properly analysed*

Richard Dawkins
Unweaving the rainbow

A *Candida albicans* PeptideAtlas

Vital Vialas, Zhi Sun, Carla Verónica Loureiro y Penha, Montserrat Carrascal,
Joaquín Abián, Lucía Monteoliva, Eric W. Deutsch, Ruedi Aebersold, Robert
L. Moritz, Concha Gil

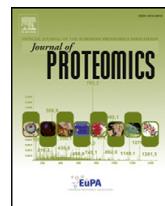
Journal of Proteomics 2014, 97, 62-68



Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



A *Candida albicans* PeptideAtlas[☆]

Vital Vialas^{a,b,*}, Zhi Sun^c, Carla Verónica Loureiro y Penha^{a,b}, Montserrat Carrascal^d, Joaquín Abián^d, Lucía Monteoliva^{a,b}, Eric W. Deutsch^c, Ruedi Aebersold^{e,f}, Robert L. Moritz^c, Concha Gil^{a,b,*}

^aDept. Microbiología II, Universidad Complutense de Madrid, Madrid, Spain^bIRYCIS: Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain^cInstitute for Systems Biology, Seattle, WA, USA^dCSIC/UAB Proteomics Laboratory, Instituto de Investigaciones Biomédicas de Barcelona—Consejo Superior de Investigaciones Científicas, Spain^eDepartment of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland^fFaculty of Science, University of Zurich, Zurich, Switzerland

ARTICLE INFO

Available online 26 June 2013

Keywords:

Candida albicans

PeptideAtlas

Proteotypic peptides

ABSTRACT

Candida albicans public proteomic datasets, though growing steadily in the last few years, still have a very limited presence in online repositories. We report here the creation of a *C. albicans* PeptideAtlas comprising near 22,000 distinct peptides at a 0.24% False Discovery Rate (FDR) that account for over 2500 canonical proteins at a 1.2% FDR. Based on data from 16 experiments, we attained coverage of 41% of the *C. albicans* open reading frame sequences (ORFs) in the database used for the searches. This PeptideAtlas provides several useful features, including comprehensive protein and peptide-centered search capabilities and visualization tools that establish a solid basis for the study of basic biological mechanisms key to virulence and pathogenesis such as dimorphism, adherence, and apoptosis. Further, it is a valuable resource for the selection of candidate proteotypic peptides for targeted proteomic experiments via Selected Reaction Monitoring (SRM) or SWATH-MS.

Biological significance

This *C. albicans* PeptideAtlas resolves the previous absence of fungal pathogens in the PeptideAtlas project. It represents the most extensive characterization of the proteome of this fungus that exists up to the current date, including evidence for uncharacterized ORFs. Through its web interface, PeptideAtlas supports the study of interesting proteins related to basic biological mechanisms key to virulence such as apoptosis, dimorphism and adherence. It also provides a valuable resource to select candidate proteotypic peptides for future (SRM) targeted proteomic experiments.

This article is part of a Special Issue entitled: Trends in Microbial Proteomics.

© 2013 Elsevier B.V. All rights reserved.

Abbreviations: SRM, Selected Reaction Monitoring; CGD, *Candida* Genome Database; FDR, False Discovery Rate; PSM, Peptide–Spectrum Match; PRIDE, Protein Identifications Database; PSS, Predicted Suitability Score; ESS, Empirical Suitability Score

[☆] This article is part of a Special Issue entitled: Trends in Microbial Proteomics.

* Corresponding authors at: Departamento de Microbiología II, Facultad de Farmacia, Universidad Complutense de Madrid, Plaza Ramón y Cajal s/n, 28040 Madrid, Spain. Tel.: +34 91 394 17 55; fax: +34 91 394 17 45.

E-mail addresses: vivialas@ucm.es (V. Vialas), conchagil@ucm.es (C. Gil).

1. Introduction

Candida albicans is a fungus of great clinical importance. In addition to asymptotically colonizing mucous membranes as a commensal in a large percentage of the population, it may cause severe opportunistic infections in specific cases such as patients with weakened immune defenses, a common circumstance in cancer and AIDS patients. *C. albicans* infections are also a threat to patients in post-surgical situations and intensive care unit stays. In this respect, invasive candidiasis remains nowadays one of the major types of nosocomial infections and a challenge in terms of economical and health costs [1–3]. From the perspective of proteomics, recent studies have provided new insights into the *C. albicans* biology and suggested new clinical biomarker candidates for diagnosis and prognosis of invasive candidiasis [4–7].

However, the clinical relevance of this organism is not reflected in the number of large-scale publicly available proteomics resources. Up to the current date, the PRIDE [8] database includes only 15 experiments accounting for 1786 identified proteins. The more *C. albicans*-focused Proteopathogen database [9] comprises several hundred protein identifications including data from gel based proteomics, and other major proteomic online resources such as the Global Proteome Machine Database (GPMDB [10]) or Tranche [11] contain no *C. albicans* data whatsoever.

As for the genomic data, according to *Candida* Genome Database (CGD), currently the most comprehensively annotated *C. albicans* sequence repository [12], the *C. albicans* genome contains 6215 ORFs (as of May 28, 2013), out of which 1497 are annotated as verified, i.e. representing genes for which there is empirical evidence that the ORF actually encodes a functionally characterized protein. In contrast, 4566 ORFs are termed uncharacterized, indicating that there exists no conclusive evidence for the existence of a protein product. This data implies that most part of the predicted proteome, over 70% of the ORFs, is still unknown or has not been properly annotated yet. An extensive characterization of the *C. albicans* proteome will therefore be of great value to increase our knowledge in proteins involved in mechanisms of virulence and infection and, thus

serves as a basis to design strategies for diagnosis, vaccination and treatment of invasive candidiasis.

Since its inception, the PeptideAtlas project [13] has encouraged mass spectrometry data submission by the community and has thus grown to a large compilation of atlases of different species including human tissue and body fluid specific builds (brain, plasma [14] and urine), microbial builds (*Halobacterium* [15], *Mycobacterium tuberculosis* [16], *Streptococcus* [17], *Leptospira*, *Plasmodium* [18], *Saccharomyces* [19] and *Schizosaccharomyces* [20]); invertebrate builds (*Caenorhabditis elegans*, *Drosophila* [21] and *Apis mellifera* [22]); and a pig and a bovine milk [23] builds. The PeptideAtlas project, as a multi-species compendium of proteomes, is continuously increasing its biological diversity. The recent *Schizosaccharomyces pombe* atlas [23] attains a large coverage of its proteome by *ad hoc* extensive fractionation and high-resolution LC-MS/MS, and contributes in the sense that some of the fission yeast biological processes have a high degree of conservation with the corresponding pathways in mammalian cells. The incorporation of *C. albicans* resolves the previous absence of fungal pathogens in the PeptideAtlas and their under representation in any public proteomic data repository.

Furthermore, the proven utility of PeptideAtlas as a resource for selecting proteotypic peptides for Selected Reaction Monitoring (SRM) [24] or SWATH-MS [25] will enable a starting point for future targeted proteomics workflows in *C. albicans*.

2. Material and methods

2.1. Empirical data compilation

Large amounts of mass spectrometry data corresponding to many and diverse measurements of the *C. albicans* proteome initially intended for different purposes were assembled in order to build the PeptideAtlas. A range of proteomic methods, protocols and different biological conditions were used to generate the data as shown in Table 1. These include membrane protein extractions [26], morphological yeast to hypha transition experiments [27] and phosphoprotein enrichment treatments. The combination of these diverse datasets resulted in an

Table 1 – List of experiments collected to construct the *C. albicans* PeptideAtlas.

# experiment	Sample (as named in the web interface)	Labeling/treatment	Instrument type	# raw files
1	Calb_acidic_subproteome	–	LTQ	3
2	Calb_memb	–	LTQ	8
3	SILAC_phos_OrbitrapVelos_1	SILAC, IMAC + TiO2	Orbitrap Velos	3
4	SILAC_phos_OrbitrapVelos_2	SILAC, IMAC + TiO2	Orbitrap Velos	3
5	SILAC_phos_OrbitrapVelos_3	SILAC, IMAC + TiO2	Orbitrap Velos	3
6	SILAC_phos_OrbitrapVelos_4	SILAC, IMAC + TiO2	Orbitrap Velos	3
7	SILAC_phos_OrbitrapXL_1A	SILAC, IMAC	Orbitrap XL	11
8	SILAC_phos_OrbitrapXL_1A_TiO2	SILAC, IMAC + TiO2	Orbitrap XL	5
9	SILAC_phos_OrbitrapXL_1B	SILAC, IMAC	Orbitrap XL	6
10	SILAC_phos_OrbitrapXL_1B_TiO2	SILAC, IMAC + TiO2	Orbitrap XL	6
11	SILAC_phos_OrbitrapXL_2	SILAC, IMAC	Orbitrap XL	6
12	SILAC_phos_OrbitrapXL_3	SILAC, IMAC	Orbitrap XL	6
13	SILAC_phos_OrbitrapXL_4	SILAC, IMAC	Orbitrap XL	5
14	Calb_extract_3TOF	–	Triple TOF	2
15	Hyphal_extract_OrbitrapVelos	–	Orbitrap Velos	4
16	Yeast_extract_OrbitrapVelos	–	Orbitrap Velos	4

unprecedented overall coverage of the *C. albicans* proteome. Protein samples were obtained as previously described in [27]. Briefly, cells of the clinical isolate SC5314 were grown in YPD medium for standard growth, whereas hyphal form growth was induced using either Lee medium pH 6.7 or heat-inactivated fetal bovine serum. Protein extracts were then obtained by mechanical cell disruption using either glass beads in the MSK cell homogenizer or the Fast-Prep cell breaker. Protein digests were obtained by trypsinization and separated via HPLC. All spectra acquisition runs were performed by LC-MS/MS in a data-dependent manner in different instruments and setups. Table 1 provides an overview of the experiments along with the instruments used for the mass spectrometry and the corresponding number of raw spectra data files that were acquired.

In addition, raw MS data from unpublished, SILAC labeled and phosphoprotein enriched samples generated from studies focused on *Candida* interaction with host immune cells and from experiments studying the hyphal and yeast-form proteomes, were added to the collection.

2.2. Peptide and protein identification

PeptideAtlas ensures consistency and quality of the stored data by processing the raw spectra sets by the Trans-Proteomic Pipeline (TPP) [28], a suite of software tools for processing shotgun proteomic datasets. The TPP tools are run in a well-established sequential pipeline spanning steps from creating appropriate standard files to be used as input by the search engine to statistical validation of protein inference and calculation of the False Discovery Rate (FDR).

The collected raw spectra files in different proprietary file formats were converted to the standard format for mass spectrometry output data mzML [29], searched using X!Tandem [30] with the K-score algorithm plug-in [31] and the output search results were converted to the search engine-independent pepXML format [32].

The target fasta sequence file used for the search was obtained from the *Candida* Genome Database (CGD) [12] at: http://www.candidagenome.org/download/sequence/C_albicans_SC5314/Assembly21/.

Common contaminants from the common Repository of Adventitious Proteins (cRAP) were appended. Then for each of these sequences, counterpart reversed decoy sequences were appended.

PeptideProphet [33] was then run on the search results to model the distributions of correctly and incorrectly assigned Peptide-to-Spectrum Matches (PSMs). It then assigns probabilities of being correct for each PSM, yielding a sensitive and flexible approach to report results in a comparable manner. Next, iProphet [34] was used to combine additional sources of evidence including multiple identifications of the same peptide across spectra, experiments, and charge and modification states, allowing a more precise integration of evidence supporting the identification of each unique peptide sequence. ProteinProphet [35] was then run to refine iProphet probabilities by adding the information at the protein level, like the number of sibling peptides within a protein and to compute final protein level probabilities. The prophet tools together combine multiple layers of evidence and refine the model iteratively to achieve an optimal analysis of the data. Finally MAYU [36] estimated FDR at different

levels for each contributing experiment and for the entire dataset based on the PSMs to decoy proteins.

This process followed the pipeline first implemented in the construction of the human plasma PeptideAtlas described in [14] and successfully applied to other builds such as the bovine milk and mammary gland PeptideAtlas [23].

2.3. Construction of the PeptideAtlas

The PeptideAtlas building process calculates the cumulative number of identified peptide and proteins across the experiments, gathers information on protein to genome location mappings and estimates the peptides' Empirical Suitability Score and Predicted Suitability Score (ESS and PSS). The genomic mappings, since *C. albicans* is not present in the Ensembl database, which is the default PeptideAtlas uses to that purpose, were extracted from a generic feature file located at the following url: http://www.candidagenome.org/download/gff/C_albicans_SC5314/C_albicans_SC5314_version_A21-s02-m05-r10_features.gff.

An overview of how the different experiments contribute, in terms of the number of identified spectra and peptides, to the atlas build is depicted in Fig. 1.

Besides, and due to the particularly rich number of identifications in experiments aimed at the detection of phosphorylated proteins (experiments #3 to #13), a similarly processed version of the PeptideAtlas was created including in this case PTMProphet results which provide, alongside each modified residue, the probability that the post-translational modification is truly detected at that site.

3. Results and discussion

3.1. Assessment of proteome coverage and functional enrichment analysis

The assembled proteomic datasets (Table 1) were subject to uniform data processing in order to build the *C. albicans*

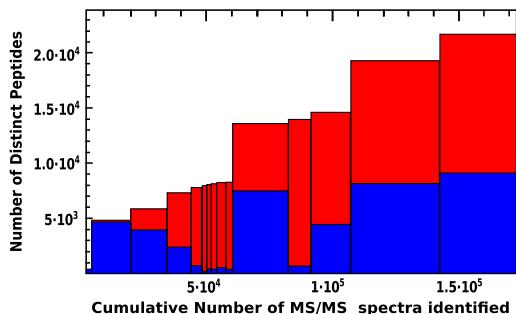


Fig. 1 – Histogram showing the cumulative number of distinct peptides in the *C. albicans* PeptideAtlas. Each bar represents a different experiment that has contributed to the build. Bar width is proportional to the number of high confidence PSMs. Height of the blue section of the bar represents the number of distinct peptides in each experiment and total height of the bar (red plus blue sections) indicates the cumulative number of peptides. The order of experiments is the same as in Table 1.

PeptideAtlas. The PSM assignment and protein inference processes were conducted by means of the consistent and robust pipeline TPP. The prophet tools integrate various levels of information and report identification results in statistical terms so that spectrum assignments, peptide to protein mappings and protein groups are statistically validated, leading to an overall improved sensitivity for a defined FDR level. As a result the generated *C. albicans* PeptideAtlas comprises 21,938 peptides identified at a 0.24% FDR allocated to 2562 proteins at a 1.2% FDR, that is, a coverage of 41.3% of the 6209 *C. albicans* translated ORF sequences from the fasta database used for searches. While the presented instance of the *C. albicans* PeptideAtlas has reached unprecedented coverage, it does not represent a final representation of the respective proteome. Like other PeptideAtlas instances for other species, the *C. albicans* atlas will be expanded upon submission and processing of new MS data generated in ongoing projects.

To determine the biological functions encompassed by the covered part of the proteome in this PeptideAtlas a Gene Ontology (GO) annotation enrichment analysis was carried out for the list of all detected *C. albicans* canonical proteins, excluding decoy hits, using the biological process ontology and Genecodis software [37]. Predictably, it generated a diverse array of clusters heterogeneously annotated, among which the largest in number of proteins are associated with the GO terms oxidation-reduction process, cellular response to drug, pathogenesis and hyphal growth respectively (Fig. 2). The enrichment in some very generic GO terms such as oxidation-reduction process, cellular response to drug and translation supports the hypothesis that the diversity of experiments assembled to build the atlas provides a representative, unbiased subset of the *C. albicans* proteome. In contrast, the more precise groups resulting from the analysis related to pathogenesis, hyphal growth and fungal-type cell wall organization are consistent with the large contribution to the atlas by the experiment aimed at identifying proteins from

cells in hyphal form and by the profusion of these sort of annotations in the source database.

As for the set of proteins present in the fasta database used for the searches that are not covered in the PeptideAtlas, they were subject to a similar analysis and were found to be enriched in annotations related to the transmembrane transport GO term (Fig. 2). These proteins are not easily observed by LC-MS/MS techniques as previously reported [20]. Also, we observed enrichment in regulation of transcription, DNA-dependent in the undetected part of the proteome. Given the short life span and low abundance of many transcription factors it is plausible that they were not detected in the collected datasets and their under representation in proteomic data has also been reported in other proteomic studies and in PeptideAtlas instances from other species [20,38,39]. The low number of protein groups significantly associated with GO annotations in the undiscovered set is understandably due to the fact that 2460 out of 3665 of the undetected protein sequences, roughly two thirds, correspond to unnamed ORFs, meaning, that little is known about their biological function.

In addition to the groups of functionally characterized proteins, this PeptideAtlas offers solid empirical evidence for the existence of 1564 proteins, showing a ProteinProphet probability score greater than 0.9, corresponding to uncharacterized ORFs in the CGD database (i.e., one-third of all 4566 uncharacterized ORFs).

3.2. Proteins of interest. Case of use

From the clinical angle, the characterization of the *C. albicans* proteome is focused on particular subproteomes, including cell surface constituents, and the set of proteins involved in the yeast-to-hypha transition. The cell wall, as the outermost cell structure represents the contact surface with host cells and therefore gathers many antigens, virulence factors and Pathogen Associated Molecular Patterns (PAMPs) [40]. Proteins

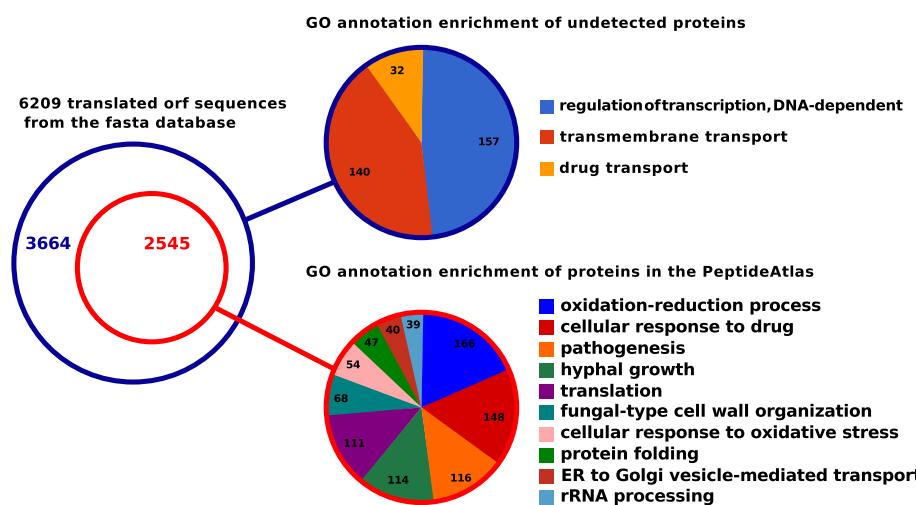


Fig. 2 – Gene Ontology annotation enrichment analysis for both the covered and undetected proteome subsets. All shown GO annotations correspond to the biological process ontology and were found significant for a p-value cut-off below 0.01.

involved in hyphal growth are also relevant in pathogenesis, in the sense that hyphae have been proven as key for invasiveness whereas the switch back to yeast form plays a role in dissemination [41].

Within these groups, a selected set of proteins of interest present in the atlas, are the adhesins from the ALS family with a role in invasiveness Als2p and Als3p; those required for cell wall biogenesis and organization glycosidases Phr1p, Phr2p and Utz2p; mannosyltransferases Pmt1p, Pmt4 and Pmt6; those involved in the cell-wall glucan metabolism Mp65p and

Ecm33p, and the hyphal cell wall constituents Hwp1, Csp37p and Rbt1p.

Other relevant proteins in the atlas are the ones related to apoptosis, since those would make an ideal target for the treatment of invasive candidiasis. Among those, the atlas contains Mcap1, Bcy1p, Ras1p and three unnamed ORFs with orthologous in other species showing roles in the apoptotic process (orf19.713, orf19.967 and orf19.7365).

For any particular proteins of interest, the PeptideAtlas web interface provides tools to explore the data. A user can

orf19.4565

Protein Name	orf19.4565
Description	BGL2 CGID:CAL0002830 COORDS:Ca2chr4_C_albicans_SC5314:445653-444727C, translated using codon table 12 (308 amino acids)
Identification status	canonical
Protein Group	orf19.4565
ProteinProphet Prob	1.00000
MultiHyp Test Prob	1.00000
Alias	BGL21, CA1541, CaO19.12034, CaO19.4565, Contig4-3104_0006, IPF1046.1, IPF22613.1,
CGID	CAL0002830, CAL0006439
Distinct Peptides	6
Total Observations	35
ProtProphet-adj NObs	15
Normalized PSMs per 100k	0.000

Distinct Observed Peptides (6)

Accession	Pre AA	Sequence	Fol AA	ESS	Best Prob	N Obs	EOS	SSRT	N Prot Map	N Gen Loc	N sample	Subpep
PAp02082372	K	DVSTFEGDLDFLK	S	1.00	1.000	15	1.00	36.83	1	1	6	0
PAp02090334	R	EDLTASELASK	I	0.74	0.999	11	0.50	22.40	1	1	3	0
PAp02099043	K	HWGVWQSDK	T	0.45	0.983	3	0.33	23.79	1	1	2	0

Predicted Highly Observable Peptides

Accession	Pre AA	Sequence	Fol AA	PSS	PSieve	ESPP	DPred	APEX	STEPP
PAp02090334	R	EDLTASELASK	I	0.91	0.93	0.70	0.65	0.05	0.73
PAp02082372	K	DVSTFEGDLDFLK	S	0.91	0.96	0.52	0.67	0.00	0.99
	K	EALQNYLPK	I	0.79	0.44	0.65	0.51	0.22	0.49

PAp02082372

Peptide Accession	PAp02082372
Peptide Sequence	DVSTFEGDLDFLK
Best Probability	1.00
Times Observed	15
Avg Molecular Weight	1484.70
pI (approx)	3.8
SSRCalc rel hydrophobicity	36.83
# Samples	6
# Protein Samples	6
Proteotypic score	1.0
# builds	4
Organisms	1

Individual Spectra

Modified Sequence	Chg	Smpl	Instr	Prob	Spectr
DVSTFEGDL ^[119] DFL ^[119] K	2	3244		2	Elu2B.27166.27166.2
DVSTFEGDLDFLK	2	3243		2	Candida_2ug.51930.51930.2
DVSTFEGDLDFLK	2	3244		2	Elu2B.27231.27231.2
⋮	⋮	⋮	⋮	⋮	⋮

DVSTFEGDLDFLK, MH⁺ 1484.7075, m/z 742.8574
File: Candida2ug-Candida_2ug.51930. Scan 51930. Precursor m/z: 743.810

* screenshots from the PeptideAtlas web interface have been adapted for clarification purposes

Fig. 3 – Protein- and peptide-centric views for Bgl2p are depicted. Distinct observed peptides are ranked by the BestProb parameter (representing the PeptideProphet probability). Of those, most probably, some will also be present in the following Predicted Highly Observable Peptides table where peptides are ranked by PSS, a combination of different prediction algorithms. For all observed peptides, spectra from the different experiments are also available.

browse through a set of protein and peptide-centric views as illustrated in Fig. 3 for the specific case of Bgl2p, a cell wall glucosyltransferase. Its corresponding observed peptides are highlighted in the protein sequence and sorted by the Empirical Suitability Score (ESS), which represents the proportion of the number of samples in which the peptide is observed with regard to the number of samples in which the original protein is observed. This parameter, in combination with others, such as a number of protein mappings, genome location and amino acid composition will help the user to select candidate proteotypic peptides for a targeted proteomics (SRM, Selected Reaction Monitoring) experiment.

Concerning those cases where a selected protein of interest is not observed in the selected build, the PeptideAtlas also provides the Predicted Suitability Score (PSS), a value resulting from the combination of different observability prediction algorithms based upon physico-chemical properties derived from the amino acid composition and previous training datasets as described in [42].

The build that assembles the phosphoprotein enrichment experiments may be of great potential interest to study biological processes such as signal transduction, since it encompasses a number of kinases and phosphatases. A total of 421 different phosphopeptides were detected and allocated to 210 phosphoproteins. The largest number of phosphorylation sites occurs in S, 410 phosphopeptides contain, at least, one phosphorylation in S; 79 phosphopeptides contain, at least, one phosphorylation in T; and 10 phosphopeptides contain one phosphorylation in Y.

4. Conclusions

This *C. albicans* PeptideAtlas build provides empirical identification evidence for 21,938 unique peptides including 421 phosphopeptides at a 0.24% peptide-level FDR that account for a high-confidence set (as defined in [14]) of 2562 canonical proteins at a 1.2% protein-level FDR representing thus a significant advance in the proteomic characterization of *C. albicans*.

Through the web interface, an important set of tools are made available to the scientific community, enabling a solid foundation to study different basic biological processes like dimorphism, signal transduction, apoptosis and the interaction with the human host. Furthermore, its value as a resource for proteotypic peptide selection is of great potential interest for future SRM experiments.

The current version of the PeptideAtlas can be found at: https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=323 and the version including PTM results at: https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=324.

Acknowledgments

The Proteomics Unit UCM-Parque Científico de Madrid is a member of the ProteoRed-Spanish National Institute for Proteomics.

We are thankful to María Luisa Hernández and Jose Antonio Reales for helping in sample obtention from the hyphal and yeast form protein extracts and to Antonio Serna for providing

the tandem mass spectra from the triple-TOF instrument. Also Aida Pitarch helped in the preparation of the manuscript.

This work was supported by BIO 2009-07654 and BIO 2012-31767 from the Ministerio de Economía y Competitividad, PROMPT (S2010/BMD-2414) from the Comunidad de Madrid, and Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Economía y Competitividad, Spanish Network for Research in Infectious Diseases (REIPI RD12/0015) -co-financed by the European Development Regional Fund "A way to achieve Europe" ERDF.

EWD, ZS, and RLM are supported in part by the National Institute of General Medical Sciences, under Grant No. R01 GM087221, 2P50 GM076547/Center for Systems Biology, the National Science Foundation MRI [Grant No. 0923536], the EU FP7 grant 'ProteomeXchange' [Grant No. 260558], and by the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg.

RA is supported in part by ERC advanced grant 'Proteomics v3.0' (Grant No. 233226) of the European Union.

R E F E R E N C E S

- [1] Wisplinghoff H, Bischoff T, Tallent SM, Seifert H, Wenzel RP, Edmond MB. Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clin Infect Dis* 2004;39:309–17.
- [2] Moran C, Grussemeyer CA, Spalding JR, Benjamin DK, Reed SD. Comparison of costs, length of stay, and mortality associated with *Candida glabrata* and *Candida albicans* bloodstream infections. *Am J Infect Control* 2010;38:78–80.
- [3] Tong KB, Murtagh KN, Lau C, Seifeldin R. The impact of esophageal candidiasis on hospital charges and costs across patient subgroups. *Curr Med Res Opin* 2008;24:167–74.
- [4] Fernández-Arenas E, Cabezón V. Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction. *Mol Cell Proteomics* 2007;6:460–78.
- [5] Pitarch A, Nombela C, Gil C. Prediction of the clinical outcome in invasive candidiasis patients based on molecular fingerprints of five anti-*Candida* antibodies in serum. *Mol Cell Proteomics* 2011;10, doi:[10.1074/mcp.M110.004010](https://doi.org/10.1074/mcp.M110.004010) [M110.004010].
- [6] Pitarch A, Nombela C, Gil C. *Candida albicans* biology and pathogenicity: insights from proteomics. *Methods Biochem Anal* 2006;49:285–330.
- [7] Pitarch A, Nombela C, Gil C. Contributions of proteomics to diagnosis, treatment, and prevention of candidiasis. *Methods Biochem Anal* 2006;49:331–61.
- [8] Vizcaíno JA, Côté RG, Csordas A, Dianes Ja, Fabregat A, Foster JM, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 2013;41:D1063–9.
- [9] Vialás V, Nogales-Cadenas R, Nombela C, Pascual-Montano A, Gil C. Proteopathogen, a protein database for studying *Candida albicans*–host interaction. *Proteomics* 2009;9:4664–8.
- [10] Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 2004;3:1234–42.
- [11] Smith BE, Hill JA, Gjukich MA, Andrews PC. Tranche distributed repository and ProteomeCommons.org. *Methods Mol Biol* 2011;696:123–45.
- [12] Costanzo MC, Arnaud MB, Skrzypek MS, Binkley G, Lane C, Miyasato SR, et al. The *Candida* Genome Database: facilitating

- research on *Candida albicans* molecular biology. *FEMS Yeast Res* 2006;6:671–84.
- [13] Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, et al. The PeptideAtlas project. *Nucleic Acids Res* 2006;34:D655–8.
- [14] Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz J A, et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics* 2011;10, doi:[10.1074/mcp.M110.006353](https://doi.org/10.1074/mcp.M110.006353) [M110.006353].
- [15] Van PT, Schmid AK, King NL, Kaur A, Pan M, Whitehead K, et al. *Halobacterium salinarum* NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage. *J Proteome Res* 2008;7:3755–64.
- [16] Schubert OT, Mouritsen J, Ludwig C, Röst HL, Rosenberger G, Arthur PK, et al. The *Mtb* Proteome Library: a resource of assays to quantify the complete proteome of *Mycobacterium tuberculosis*. *Cell Host Microbe* 2013;13:602–12.
- [17] Lange V, Malmström Ja, Didion J, King NL, Johansson BP, Schäfer J, et al. Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. *Mol Cell Proteomics* 2008;7:1489–500.
- [18] Lindner SE, Swearingen KE, Harupa A, Vaughan AM, Sinnis P, Moritz RL, et al. Total and putative surface proteomics of malaria parasite salivary gland sporozoites. *Mol Cell Proteomics* 2013;12, doi:[10.1074/mcp.M112.024505](https://doi.org/10.1074/mcp.M112.024505) [M110.024505].
- [19] King NL, Deutsch EW, Ranish JA, Nesvizhskii AI, Eddes JS, Mallick P, et al. Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. *Genome Biol* 2006;7:R106.
- [20] Gunaratne J, Schmidt A, Quandt A, Neo SP, Sarac OS, Gracia T, et al. Extensive mass spectrometry-based analysis of the fission yeast proteome: the *S. pombe* PeptideAtlas. *Mol Cell Proteomics* 2013;12, doi:[10.1074/mcp.M112.023754](https://doi.org/10.1074/mcp.M112.023754) [M112.023754].
- [21] Loevenich SN, Brunner E, King NL, Deutsch EW, Stein SE, Aebersold R, et al. The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. *BMC Bioinformatics* 2009;10:59.
- [22] Chan QWT, Parker R, Sun Z, Deutsch EW, Foster LJ. A honey bee (*Apis mellifera* L.) PeptideAtlas crossing castes and tissues. *BMC Genomics* 2011;12:290.
- [23] Bislev S, Deutsch E, Sun Z. A bovine PeptideAtlas of milk and mammary gland proteomes. *Proteomics* 2012;12:2895–9.
- [24] Deutsch E, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 2008;9:429–34.
- [25] Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* 2012;11, doi:[10.1074/mcp.O111.016717](https://doi.org/10.1074/mcp.O111.016717) [O111.016717].
- [26] Cabezón V, Llama-Palacios A, Nombela C, Monteoliva L, Gil C. Analysis of *Candida albicans* plasma membrane proteome. *Proteomics* 2009;9:4770–86.
- [27] Monteoliva L, Martinez-Lopez R. Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition. *J Proteome Res* 2010;10:502–17.
- [28] Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 2010;10:1150–9.
- [29] Martens L, Chambers M, Sturm M. mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011;10, doi:[10.1074/mcp.R110.000133](https://doi.org/10.1074/mcp.R110.000133) [R110.000133].
- [30] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–7.
- [31] MacLean B, Eng J, Beavis R, McIntosh M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* 2006;22:2830–2.
- [32] Keller A, Eng J, Zhang N. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 2005;1 [2005.0017].
- [33] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;74:5383–92.
- [34] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* 2011;10, doi:[10.1074/mcp.M111.007690](https://doi.org/10.1074/mcp.M111.007690) [M111.007690].
- [35] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;75:4646–58.
- [36] Reiter L, Claassen M, Schrimpf S. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* 2009;8:2405–17.
- [37] Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res* 2012;40:W478–83.
- [38] Ding C, Chan DW, Liu W, Liu M, Li D, Song L, et al. Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements. *Proc Natl Acad Sci U S A* 2013;110:6771–6.
- [39] Simicevic J, Schmid AW, Gilardoni PA, Zoller B, Raghav SK, Krier I, et al. Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nat Methods* 2013;10:570–6.
- [40] Vialás V, Perumal P, Gutierrez D, Ximénez-Eembún P, Nombela C, Gil C, et al. Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells. *Proteomics* 2012;8:2331–9.
- [41] Saville SP, Lazzell AL, Monteagudo C, Lopez-Ribot JL. Engineered control of cell morphology *in vivo* reveals distinct roles for yeast and filamentous forms of *Candida albicans* during infection. *Eukaryot Cell* 2003;2:1053–60.
- [42] Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* 2007;25:125–31.

***Subcellular fractionation and different growing conditions lead
to a large increase in the proteome coverage in the Candida
albicans PeptideAtlas***

DISCUSIÓN

Discusión

El abordaje de experimentos proteómicos usando *Candida albicans* como organismo modelo para el estudio de infecciones fúngicas del tipo de las candidiasis ha permitido profundizar en algunos de los aspectos más básicos de su biología y también en algunos procesos y mecanismos interesantes desde el punto de vista clínico. Así, se han abordado estudios que intentan identificar la mayor cantidad de proteínas posibles en extractos celulares totales, también ensayos que estudian subproteomas como el implicado en la transición levadura a hifa, enfocados a las proteínas de la superficie celular y secretadas, o a las implicadas en la formación de biopelículas entre otros aspectos. Esta heterogenidad existente en los diseños experimentales también se puede observar en la forma en que el análisis continúa trás la adquisición de espectros de masas por LC-MS/MS. Los datos son analizados bioinformáticamente utilizando distintos motores de búsqueda, distintos métodos de validación, flujos de análisis diferentes en definitiva. Además, hasta hace relativamente poco tiempo, los primeros años del siglo XXI, la presencia de resultados de proteómica en repositorios públicos era muy escasa, casi nula, y en ocasiones poco fiable.

En este contexto las bases de datos y estándares desarrollados en Proteómica tienen un papel esencial para analizar, compartir y difundir resultados. Las herramientas aquí descritas contribuyen a estos objetivos.

DISCUSIÓN

Desarrollo de una aplicación web para datos de proteómica a gran escala de *Candida albicans*

La base de datos y herramienta web Proteopathogen es la primera aplicación *on line* descrita para recoger y analizar resultados de Proteómica centrados en el estudio de hongos patógenos usando principalmente el modelo de *Candida albicans*.

Durante algunos años Proteopathogen ha recogido datos experimentales de experimentos de *Candida albicans* ([Cabezón et al., 2009](#); [Monteoliva et al., 2011](#); [Vialás et al., 2012](#)) y otras especies como *Aspergillus fumigatus* () facilitando la visualización y el análisis de los resultados a los usuarios del laboratorio, pero también su examen por parte de los revisores de las revistas científicas del campo de la Proteómica.

En ese tiempo, la Iniciativa de Estandarización del Proyecto Proteoma Humano HUPO-PSI (*Human Proteome Organization - Proteomics Standards Initiative*) ha desarrollado y promovido el uso de estándares en Proteómica, formatos que faciliten el intercambio, re-análisis y comparación de protocolos experimentales, datos y resultados entre distintos laboratorios. En este sentido Proteopathogen se ha beneficiado de la aparición del estándar mzIdentML ([Jones et al., 2012](#)), el formato creado por HUPO-PSI y posteriormente adoptado por la comunidad, para recoger la información relacionada con la identificación de péptidos y proteínas, es decir el análisis bioinformático desde la asignación de los PSM hasta la presentación de una lista de proteínas.

Proteopathogen emplea el lenguaje de programación interpretado Ruby. A diferencia de lenguajes compilados (Java, C/C++) que se usan frecuentemente en otras herramientas para la visualización de contenidos en formatos de proteómica ([Griss et al., 2012](#); [Barsnes et al., 2011](#)), esto posibilita una manera muy rápida y flexible de implementar un sistema de extracción de la información de archivos XML. La plataforma de desarrollo web Rails, también basada en Ruby, además cuenta con un gran soporte por parte de la comunidad informática y se está convirtiendo en una de las tecnologías de referencia elegida por los programadores de aplicaciones web.

Con la adopción del formato mzIdentML (version 1.1.0) como fuente de resul-

DISCUSIÓN

tados, Proteopathogen ha adquirido la capacidad de crecer de manera robusta y fiable. Por una parte, usar un único tipo de formato, estándar, como fuente de datos ha permitido desarrollar un *software* estable que extrae la información independientemente de cómo se hayan obtenido los resultados. Pero además, el uso de este estándar permite realizar validaciones, un control de calidad, tanto de la estructura, sintaxis y orden de los elementos XML de los archivos (validación semántica), como del contenido mínimo (validación MIAPE) usando para ello algunas herramientas como la creada por HUPO-PSI, mzidValidator ([Ghali et al., 2013](#)).

La base de datos relacional implementada *ad hoc* para recoger el contenido de los archivos constituye la base fundamental de la aplicación y permite que Proteopathogen sea el primer recurso *on-line* descrito que recoge y permite visualizar resultados, individualmente para cada archivo mzIdentML o en conjunto, procedentes de múltiples experimentos en el campo de los hongos patógenos usando principalmente el modelo de *C. albicans*.

En cuanto al futuro de esta herramienta bioinformática, es importante destacar que el desarrollo de nuevos formatos estándar es continuo. Si bien no es probable que mzIdentML sea sustituido por otro, sí es cierto que pronto verá una versión actualizada (1.2.0). En ese escenario, los programas que transfieren el contenido a la base de datos deberán ser adaptados, aunque previsiblemente los cambios no provocarán incompatibilidades sino que serán fundamentalmente aditivos añadiendo algún nuevo tipo de información. Para ello, la flexibilidad que proporciona el uso del lenguaje de programación Ruby y el entorno de desarrollo web Rails permitirá que los cambios desde el periodo de pruebas hasta la producción en el servidor puedan implementarse rápidamente y con seguridad.

Otro tipo de posible mejora en la aplicación podría consistir en implementar un nuevo modo de leer los archivos mzIdentML. Esta posibilidad vendría motivada por la creciente capacidad de adquisición de datos de los espectrómetros de masas y las mejoras en el análisis bioinformático subsiguiente, que se traduce en archivos de resultados que llegan a tener un gran tamaño para archivos de texto (hasta el orden de gigabytes). El modo en que Proteopathogen lee los archivos consiste en una representación en memoria de toda la jerarquía xml (*parser DOM*). Esto, que es muy efectivo para localizar los distintos elementos referenciados en el contenido

DISCUSIÓN

y guardarlos en cada tabla correspondiente de la base de datos, puede convertirse en una tarea difícil (requerir computadores con gran memoria de trabajo, RAM) o incluso imposible en algunos casos. Por este motivo puede ser interesante explorar otro tipo de implementaciones de lectura de archivos XML (*parser SAX*) que permitan leer la información contenida en archivos de gran tamaño.

Desarrollo de un PeptideAtlas de *Candida albicans*

El proyecto PeptideAtlas, desde su inicio hace casi una década (Desiere et al., 2006), ha animado a la comunidad proteómica a contribuir con sus experimentos y resultados de LC-MS/MS. La particularidad de PeptideAtlas reside en que, a diferencia de otros grandes repositorios públicos de proteómica como PRIDE, todos los resultados son analizados mediante un flujo de trabajo homogéneo, proporcionado por las herramientas de *software* agrupadas en TPP. Así, el proyecto se ha convertido en un gran compendio de atlas de diferentes especies caracterizados por una reconocida calidad y fiabilidad en las identificaciones de péptidos y proteínas.

La creación del PeptideAtlas de *Candida albicans* supuso la incorporación por primera vez de un modelo de hongo patógeno en el proyecto PeptideAtlas y la primera recopilación de resultados de proteómica que alcanzó una gran escala para este organismo. En su primera versión (Vialas et al., 2013), se alcanzó una cobertura del proteoma sin precedentes para resultados experimentales agrupados en un solo proyecto, un PeptideAtlas para *Candida albicans*. Casi 22.000 péptidos correspondientes a más de 2500 proteínas suponían una cobertura de más del 40 % del proteoma predicho.

Pero PeptideAtlas permite que los resultados sean reprocesados cuando se obtienen nuevos resultados de LC-MS/MS o cuando se desarrollan nuevas mejoras en los motores de búsqueda o en el *software* de análisis. Así, trás el PeptideAtlas original se obtuvieron nuevos resultados experimentales, bien específicamente destinados a la ampliación de la cobertura del proteoma mediante fraccionamientos extensivos a varios niveles: subcelular (centrifugaciones), proteína (SDS-PAGE) y péptido (OFF-GEL); o bien procedentes de trabajos destinados al estudio de proteínas de la pared celular () o secretadas mediante vesículas (Gil-Bona et al., 2014). Ade-

DISCUSIÓN

más en ese tiempo apareció un nuevo ensamblaje de la secuencia de *C. albicans* que por primera vez incluía secuencias específicas de alelo. (Muzzey et al., 2013). En este contexto, con los nuevos sets de datos y la nueva información de secuencia disponible, el PeptideAtlas original ha sido reanalizado, en conjunto con los sets de datos que formaban la primera versión, y empleando tres motores de búsqueda (Sequest, OMSSA y Comet) para finalmente obtener una cobertura de dos terceras partes del proteoma predicho de *C. albicans*. Más de 71.000 péptidos asignados a 4174 secuencias de proteínas (para un FDR de 0.10 % a nivel de PSM) suponen exactamente 66.17 % del proteoma, y con respecto a la versión inicial, un incremento de 3 veces el número de péptidos y 1.6 veces el de proteínas.

Con este notable incremento, este PeptideAtlas continúa siendo el recurso público de datos proteómicos de *C. albicans* más exhaustivo.

Además de describir una lista de proteínas que representa un 66 % del proteoma predicho en CGD (*Candida Genome Database*), el PeptideAtlas de *C. albicans* proporciona evidencia empírica sólida de la existencia de proteínas para dos terceras partes de los genes que en CGD se denominan *uncharacterized* por carecer de un producto proteico caracterizado.

Por último, y para favorecer la comunicación e interconexión entre ambos recursos, CGD y PeptideAtlas, se ha contactado con los desarrolladores e impulsores de CGD proporcionándoles un formato de enlace para que a través de la información en la pestaña *proteína* en CGD, se pueda acceder a los datos correspondientes a la identificación por MS en PeptideAtlas.

CONCLUSIONES

Conclusiones

1. La base de datos y aplicación web Proteopathogen ha demostrado ser una herramienta de gran utilidad para la visualización y análisis de resultados de proteómica en experimentos que usan *Candida albicans* como organismo modelo de estudio de hongos patógenos.
2. La adopción del estándar mzIdentML como formato de origen para incorporar nuevos datos en Proteopathogen asegura la estabilidad y futuro de este proyecto facilitando la incorporación de resultados procedentes de nuevos experimentos.
3. El PeptideAtlas de *Candida albicans* desarrollado supone la caracterización más exhaustiva del proteoma de este organismo. y es el recurso más completo y fiable disponible públicamente.

Bibliografía

*Y así, del mucho leer y del poco dormir, se le
secó el celebro de manera que vino a perder
el juicio.*

*Primera parte de El Ingenioso Caballero Don
Quijote de la Mancha
Miguel de Cervantes Saavedra*

ABDI, H. H. The Bonferroni and Sidak Corrections for Multiple Comparisons. *Encyclopedia of Measurement and Statistics*, vol. 1, páginas 1–9, 2007.

BARSNES, H., VAUDEL, M., COLAERT, N., HELSENS, K., SICKMANN, A., BERVEN, F. S. y MARTENS, L. compomics-utilities: an open-source Java library for computational proteomics. *BMC bioinformatics*, vol. 12, página 70, 2011. ISSN 1471-2105.

BENJAMINI, Y. y HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57(1), páginas 289–300, 1995. ISSN 00359246.

CABEZÓN, V., LLAMA-PALACIOS, A., NOMBELA, C., MONTEOLIVA, L. y GIL, C. Analysis of *Candida albicans* plasma membrane proteome. *Proteomics*, vol. 9(20), páginas 4770–86, 2009. ISSN 1615-9861.

CALDERONE, R. A. y FONZI, W. A. Virulence factors of *Candida albicans*. *Trends in Microbiology*, vol. 9(7), páginas 327–335, 2001. ISSN 0966842X.

BIBLIOGRAFÍA

- CHOI, H. y NESVIZHSKII, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of proteome research*, vol. 7(1), páginas 254–65, 2008. ISSN 1535-3893.
- DESIERE, F., DEUTSCH, E. W., KING, N. L., NESVIZHSKII, A. I., MALLICK, P., ENG, J., CHEN, S., EDDES, J., LOEVENICH, S. N. y AEBERSOLD, R. The PeptideAtlas project. *Nucleic acids research*, vol. 34(Database issue), páginas D655–8, 2006. ISSN 1362-4962.
- DEUTSCH, E. E. W., MENDOZA, L., SHTEYNBERG, D., FARRAH, T., LAM, H., TASMAN, N., SUN, Z., NILSSON, E., PRATT, B., PRAZEN, B., ENG, J. K., MARTIN, D. B., NESVIZHSKII, A. I. y AEBERSOLD, R. A guided tour of the TransProteomic Pipeline. *Proteomics*, vol. 10(6), páginas 1150–1159, 2010. ISSN 1615-9861.
- DEUTSCH, E. W. Mass spectrometer output file format mzML. *Methods in molecular biology (Clifton, N.J.)*, vol. 604, páginas 319–31, 2010. ISSN 1940-6029.
- DEUTSCH, E. W., CHAMBERS, M., NEUMANN, S., LEVANDER, F., BINZ, P.-A., SHOFSTAHL, J., CAMPBELL, D. S., MENDOZA, L., OVELLEIRO, D., HELSENS, K., MARTENS, L., AEBERSOLD, R., MORITZ, R. L. y BRUSNIAK, M.-Y. TraML—a standard format for exchange of selected reaction monitoring transition lists. *Molecular & cellular proteomics : MCP*, vol. 11(4), página R111.015040, 2012. ISSN 1535-9484.
- DING, Y., CHOI, H. y NESVIZHSKII, A. I. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. *Journal of proteome research*, vol. 7(11), páginas 4878–89, 2008. ISSN 1535-3893.
- ELIAS, J. E. y GYGI, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, vol. 4(3), páginas 207–14, 2007. ISSN 1548-7091.
- ENG, J. K., JAHAN, T. A. y HOOPMANN, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics*, vol. 13(1), páginas 22–4, 2013. ISSN 1615-9861.

BIBLIOGRAFÍA

- ENG, J. K., McCORMACK, A. L. y YATES, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, vol. 5(11), páginas 976–89, 1994. ISSN 1044-0305.
- FENN, J. B., MANN, M., MENG, C. K., WONG, S. F. y WHITEHOUSE, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science (New York, N.Y.)*, vol. 246(4926), páginas 64–71, 1989. ISSN 0036-8075.
- FENYÖ, D. y BEAVIS, R. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry*, vol. 75(4), páginas 768–774, 2003. ISSN 0003-2700.
- GEER, L. Y., MARKEY, S. P., KOWALAK, J. A., WAGNER, L., XU, M., MAYNARD, D. M., YANG, X., SHI, W. y BRYANT, S. H. Open mass spectrometry search algorithm. *Journal of proteome research*, vol. 3(5), páginas 958–64, ???? ISSN 1535-3893.
- GERBER, S. A., RUSH, J., STEMMAN, O., KIRSCHNER, M. W. y GYGI, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100(12), páginas 6940–5, 2003. ISSN 0027-8424.
- GHALI, F., KRISHNA, R., LUKASSE, P., MARTÍNEZ-BARTOLOMÉ, S., REISINGER, F., HERMJAKOB, H., VIZCAÍNO, J. A. y JONES, A. R. Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML. *Molecular & cellular proteomics : MCP*, vol. 12(11), páginas 3026–35, 2013. ISSN 1535-9484.
- GIL-BONA, A., LLAMA-PALACIOS, A., PARRA, C. M., VIVANCO, F., NOMBELA, C., MONTEOLIVA, L. y GIL, C. Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in *Candida albicans*. *Journal of proteome research*, 2014. ISSN 1535-3907.
- GRISS, J., REISINGER, F., HERMJAKOB, H. y VIZCAÍNO, J. A. jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spec-

BIBLIOGRAFÍA

- trometry data formats. *Proteomics*, vol. 12(6), páginas 795–8, 2012. ISSN 1615-9861.
- HATT, P. D., QUADRONI, M., STAUDENMANN, W. y JAMES, P. Concentration of, and SDS Removal from Proteins Isolated from Multiple two-Dimensional Electrophoresis Gels. *European Journal of Biochemistry*, vol. 246(2), páginas 336–343, 1997. ISSN 0014-2956.
- JONES, A. R., EISENACHER, M., MAYER, G., KOHLBACHER, O., SIEPEN, J., HUBBARD, S. J., SELLEY, J. N., SEARLE, B. C., SHOFSTAHL, J., SEYMOUR, S. L., JULIAN, R., BINZ, P.-A., DEUTSCH, E. W., HERMJAKOB, H., REISINGER, F., GRISS, J., VIZCAÍNO, J. A., CHAMBERS, M., PIZARRO, A. y CREASY, D. The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular & cellular proteomics : MCP*, vol. 11(7), página M111.014381, 2012. ISSN 1535-9484.
- KÄLL, L., STOREY, J. D., MACCOSS, M. J. y NOBLE, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, vol. 7(1), páginas 29–34, 2008. ISSN 1535-3893.
- KARAS, M. y HILLENKAMP, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry*, vol. 60(20), páginas 2299–301, 1988. ISSN 0003-2700.
- KELLER, A., NESVIZHSKII, A. I., KOLKER, E. y AEBERSOLD, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry*, vol. 74(20), páginas 5383–92, 2002. ISSN 0003-2700.
- KUHN, E., WU, J., KARL, J., LIAO, H., ZOLG, W. y GUILD, B. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and ¹³C-labeled peptide standards. *Proteomics*, vol. 4(4), páginas 1175–86, 2004. ISSN 1615-9853.
- LAM, H., DEUTSCH, E. W., EDDES, J. S., ENG, J. K., KING, N., STEIN, S. E. y AEBERSOLD, R. Development and validation of a spectral library searching

BIBLIOGRAFÍA

- method for peptide identification from MS/MS. *Proteomics*, vol. 7(5), páginas 655–67, 2007. ISSN 1615-9853.
- MA, K., VITEK, O. y NESVIZHSKII, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC bioinformatics*, vol. 13 Suppl 1(Suppl 16), página S1, 2012. ISSN 1471-2105.
- MACLEAN, B., ENG, J., BEAVIS, R. y MCINTOSH, M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics*, vol. 22(22), páginas 2830–2832, 2006. ISSN 1367-4803, 1460-2059.
- MAKAROV, A. Electrostatic axially harmonic orbital trapping, a high performance technique of mass analysis. *Analytical chemistry*, vol. 72(6), páginas 1156–62, 2000. ISSN 1520-6882.
- MONTEOLIVA, L., MARTINEZ-LOPEZ, R., PITARCH, A., HERNAEZ, M. L., SERNA, A., NOMBELA, C., ALBAR, J. P. y GIL, C. Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition. *Journal of Proteome Research*, vol. 10(2), páginas 502–517, 2011. ISSN 15353893.
- MUZZEY, D., SCHWARTZ, K., WEISSMAN, J. S. y SHERLOCK, G. Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome biology*, vol. 14(9), página R97, 2013. ISSN 1465-6914.
- NAVARRO, P. y VÁZQUEZ, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *Journal of proteome research*, vol. 8(4), páginas 1792–6, 2009. ISSN 1535-3893.
- NESVIZHSKII, A. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, vol. 73(11), páginas 2092–2123, 2010. ISSN 1876-7737.
- NESVIZHSKII, A. I. Protein identification by tandem mass spectrometry and sequence database searching. *Methods in molecular biology (Clifton, N.J.)*, vol. 367, páginas 87–119, 2007. ISSN 1064-3745.

BIBLIOGRAFÍA

- NESVIZHSKII, A. I. y AEBERSOLD, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & Cellular Proteomics*, vol. 4(10), páginas 1419–40, 2005. ISSN 1535-9476.
- NESVIZHSKII, A. I., KELLER, A., KOLKER, E. y AEBERSOLD, R. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, vol. 75(17), páginas 4646–58, 2003. ISSN 0003-2700.
- OLSEN, J. V., MACEK, B., LANGE, O., MAKAROV, A., HORNING, S. y MANN, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods*, vol. 4(9), páginas 709–12, 2007. ISSN 1548-7091.
- REITER, L., CLAASSEN, M., SCHRIMPFF, S. S. P., JOVANOVIC, M., SCHMIDT, A., BUHMANN, J. M., HENGARTNER, M. O. y AEBERSOLD, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & Cellular Proteomics*, vol. 8(11), páginas 2405–2417, 2009. ISSN 1535-9484.
- REITER, L., RINNER, O., PICOTTI, P., HÜTTENHAIN, R., BECK, M., BRUSNIAK, M.-Y., HENGARTNER, M. O. y AEBERSOLD, R. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature methods*, vol. 8(5), páginas 430–5, 2011. ISSN 1548-7105.
- ROEPSTORFF, P. y FOHLMAN, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry*, vol. 11(11), página 601, 1984. ISSN 0306-042X.
- ROGOWSKA-WRZESINSKA, A., LE BIHAN, M.-C., THAYSEN-ANDERSEN, M. y ROEPSTORFF, P. 2D gels still have a niche in proteomics. *Journal of proteomics*, vol. 88, páginas 4–13, 2013. ISSN 1876-7737.
- SHTEYNBERG, D., DEUTSCH, E. W., LAM, H., ENG, J. K., SUN, Z., TASMAN, N., MENDOZA, L., MORITZ, R. L., AEBERSOLD, R. y NESVIZHSKII, A. I. iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Molecular & Cellular Proteomics*, vol. 10(12), páginas M111.007690–M111.007690, 2011. ISSN 1535-9476.

BIBLIOGRAFÍA

- STEEN, H. y MANN, M. The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, vol. 5(9), páginas 699–711, 2004. ISSN 1471-0072.
- SYKA, J. E. P., COON, J. J., SCHROEDER, M. J., SHABANOWITZ, J. y HUNT, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101(26), páginas 9528–33, 2004. ISSN 0027-8424.
- TANAKA, K., WAKI, H., IDO, Y., AKITA, S., YOSHIDA, Y., YOSHIDA, T. y MATSUO, T. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, vol. 2(8), páginas 151–153, 1988. ISSN 0951-4198.
- VIALÁS, V., PERUMAL, P., GUTIERREZ, D., XIMÉNEZ-EMBÚN, P., NOMBELA, C., GIL, C. y CHAFFIN, W. L. Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells. *Proteomics*, 2012. ISSN 16159861.
- VIALAS, V., SUN, Z., LOUREIRO Y PENHA, C. V., CARRASCAL, M., ABIÁN, J., MONTEOLIVA, L., DEUTSCH, E. W., AEBERSOLD, R., MORITZ, R. L. y GIL, C. A *Candida albicans* PeptideAtlas. *Journal of proteomics*, vol. 97, páginas 62–8, 2013. ISSN 1876-7737.
- WASINGER, V. C., CORDWELL, S. J., CERPA-POLJAK, A., YAN, J. X., GOOLEY, A. A., WILKINS, M. R., DUNCAN, M. W., HARRIS, R., WILLIAMS, K. L. y HUMPHERY-SMITH, I. Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*, vol. 16(7), páginas 1090–4, 1995. ISSN 0173-0835.
- ZUBAREV, R. A., KELLEHER, N. L. y McLAFFERTY, F. W. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society*, vol. 120(13), páginas 3265–3266, 1998. ISSN 0002-7863.

List a de acrónimos

- 1D-PAGE *Monodimensional PoliAcrylamide Gel Electrophoresis*
- 2D-PAGE *Bidimensional PoliAcrylamide Gel Electrophoresis*
- CGD *Candida Genome Database*
- CID *Collision Induced Dissociation*
- DDA *Data Dependent Acquisition*
- ECD *Electron Capture Dissociation*
- EM..... *Expectation-Maximization*
- ESI *Electro-Spray Ionization*
- ETD..... *Electron Transfer Dissociation*
- FAB..... *Fast Atom Bombardment*
- FD *Field Desorption*
- FDR *False Discovery Rate*
- FNR *False Negative Rate*
- FPR..... *False Positive Rate*
- FTICR *Fourier Transform Ion Cyclotron Resonance*

BIBLIOGRAFÍA

- FWHM *Full Width at Half Mass*
- HCD *Higher Energy Collision Dissociation*
- HPLC *High Performance Liquid Chromatography*
- HUPO-PSI.... *Human Proteome Organization - Proteomics Standards Initiative*
- LIT *Linear Ion Trap*
- LTQ *Linear Trap Quadrupole*
- MALDI..... *Matrix Assisted Laser Desorption Ionization*
- MS/MS *Tandem Mass Spectrometry*
- NIST *National Institute for Standards and Technology*
- NMC *Number of Missed Cleavages*
- NRS *Number of Replicate Spectra*
- NSE *Number of Sibling Experiments*
- NSI *Number of Sibling Ions*
- NSM *Number of Sibling Modifications*
- NSP *Number of Sibling Peptides*
- NSS *Number of Sibling Searches*
- NTT..... *Number of Tryptic Termini*
- OMSSA..... *Open Mass Spectrometry Search Algorithm*
- PAGE..... *PolyAcrylamide Gel Electrophoresis*
- PD *Plasma Desorption*

BIBLIOGRAFÍA

- PMF *Peptide Mass Fingerprint*
- PRIDE *Protein Identifications Database*
- PSM *Peptide-Spectrum Match*
- PTM *Post-Translational Modification*
- QIT *Quadrupole Ion Trap*
- QTOF *Quadrupole-Time Of Flight*
- QTRAP *Quadrupole-Ion Trap*
- RP-HPLC *Reverse Phase High Performance Liquid Chromatography*
- SDS-PAGE .. *Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis*
- SLD *Soft Laser Desorption*
- TOF *Time Of Flight*
- TPP *Trans Proteomics Pipeline*
- TPR *True Positive Rate*

*—¿Qué te parece desto, Sancho? — Dijo Don Quijote —
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.*

*Segunda parte de El Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes Saavedra*

*—Buena está — dijo Sancho —; fírmela vuestra merced.
—No es menester firmarla — dijo Don Quijote—,
sino solamente poner mi rúbrica.*

*Primera parte de El Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes Saavedra*

