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

Octubre 2014

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

Octubre 2014

A mis padres

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

Carl Sagan

Agradecimientos

A todos a los que los que la Bioinformática les hace tilín.

Gracias a bla blad

Por último, gracias a la gente que ha hecho y que participa en StackOverflow, siempre hay gurús en internet que saben más que tú.

Resumen

...

...

...

Índice

Índice de figuras

Índice de Tablas

Introducción

Introducción

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

Theodosius Dobzhansky

Tradicionalmente, el gen se ha concebido como la unidad fundamental -el átomo- de la vida, sometida a la acción de la selección natural. Así definió Richard Dawkins en el Gen Egoísta al gen, la unidad indivisible auto-replicante, mientras que los individuos y sus conductas eran meras *máquinas de supervivencia*. Sin embargo, es el fenotipo y no el genotipo lo que interactúa con el ambiente y con otros organismos. Las proteínas, los *ladrillos* con que se construye la vida, sí son visibles, a diferencia de los genes, a la selección natural. Por otra parte, el clásico dogma central de la Biología Molecular caducó hace ya tiempo y hoy lo recordamos, más bien, como una sobresimplificación. Actualmente, el emergente campo de la Proteogenómica da cuenta de la intrincada red de procesos regulatorios de transferencia de información entre el gen y la proteína. La Proteómica por su parte, le debe a la Genómica el reconocimiento y agradecimiento de haber abierto camino en la Biotecnología moderna. La Bioinformática, en este panorama, tiene un papel integrador. Al igual que la Proteómica, se sirve de diferentes tecnologías que avanzan y se retroalimentan sinérgicamente. Así la Proteómica se beneficia de los avances en Espectrometría de Masas, y estos instrumentos progresan en función de la demanda en investigación. De la misma manera, la Proteómica Computacional, la parte de la Bioinformática mas cercana a la Proteómica, evoluciona para facilitar el análisis de los datos que los investigadores requieren, pero también se beneficia de la incesante, creciente capacidad de procesamiento en las computadoras actuales.

Proteómica. Conceptos generales

El concepto de Proteoma fue acuñado originalmente por Marc Wilkins en 1994 en analogía al concepto de Genoma. 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 permanece constante 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 estadío del desarrollo entre otros factores. Además existen niveles de complejidad adicional en el curso de información desde el gen a la proteína como el *splicing* alternativo y las modificaciones post-traduccionales. 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 péptidos y proteínas como la electroforesis PAGE y la cromatografía líquida permitieron que la Proteómica se consolidara y extendiera como disciplina científica.

La figura 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.

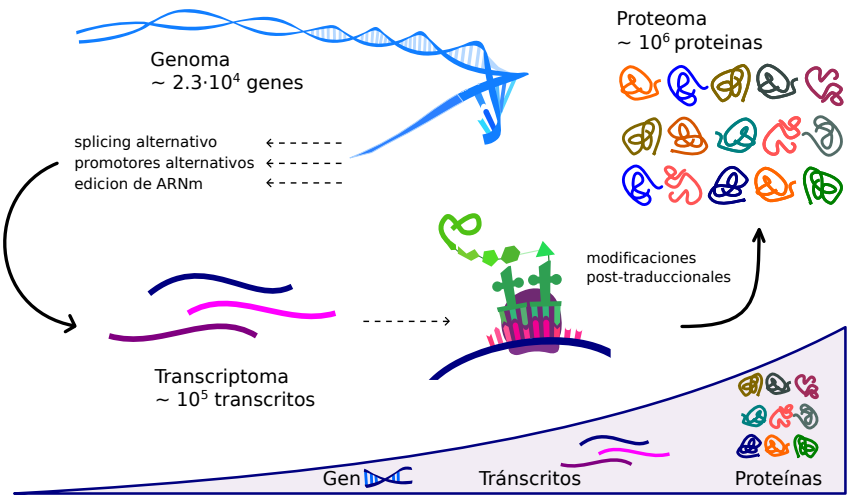


Figura 1: Aumento de complejidad desde el genoma hacia el proteoma

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 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) (**?**) y Desorción Suave por Láser (SLD) (**?**) permitieron que las grandes y frágiles moléculas biológicas como las proteínas pudieran ser ionizadas y volatilizadas 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,

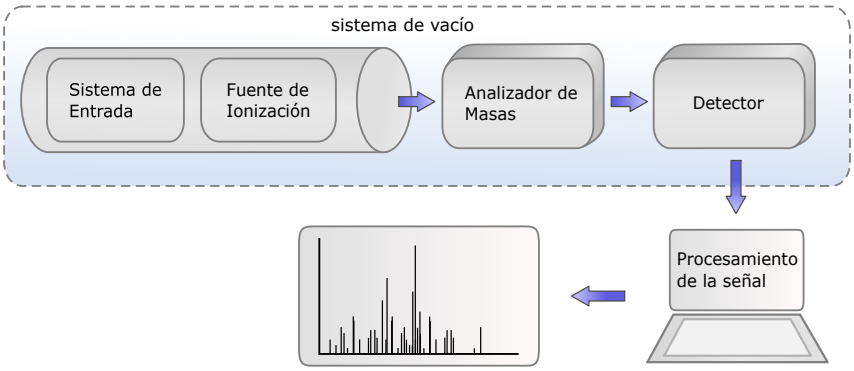


Figura 2: Esquema de un espectrómetro de masas

Alemania (éstos discutiblemente no galardonados) habían ideado una técnica similar que, en este caso, denominaron Desorción/Ionización Láser Asistida por Matriz (MALDI) (**?**) Aunque MALDI no fue aplicado 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.

Un espectrómetro de masas es, en esencia, una balanza de precisión molecular capaz de medir, hasta un determinado límite de sensiblidad, 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 físi-

ca 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 del ESI, el sistema de entrada y la fuente de iones forman parte de un único componente.

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.

   - En **ESI**, 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, adquieren una forma cónica y forman un aerosol. Se produce entonces la desolvatación por evaporación del solvente. Así, las micro-gotas del aerosol disminuyen su tamaño, 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 del solvente (explosión de Coulomb) quedando iones de analito en suspensión que son introducidos en un sistema vacío hacia el espectrómetro.
   - **MALDI** 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. Generalmente la ionización MALDI, por su carácter pulsante, se usa acoplada a analizadores de tipo Tiempo de Vuelo (TOF) que miden el tiempo que tardan los analitos ionizados en llegar al detector a través del vacío.

- **FAB**, *Fast Atom Bombardment* o Bombardeo Rápido Atómico consiste en hacer incidir un haz de alta energía de átomos de un gas inerte (Argón o Xenón) sobre el analito provocando de esa forma su ionización.
- En la ionización por **FD**, *Field Desorption* o Desorción en Campo Eléctrico la muestra se encuentra sometida a un campo eléctrico creado en una superficie, generalmente un filamento de tungsteno, llamado *emisor*. Al superar un umbral de diferencia de potencial se produce la desorción e ionización del analito.
- **PD**, *Plasma Desorption* o Desorción por Plasma consiste en hacer uso de un isótopo radiactivo ^{252}Cf que al sufrir su fisión espontánea produce dos partículas de alta energía con trayectorias opuestas, (generalmente ^{144}Cs y y ^{108}Tc) que impactan sobre la muestra provocando su desorción e ionización.

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 espectrómetros híbridos. Así, un analizador tipo *cuadrupolo* puede encontrarse acoplado con un analizador de *tiempo de vuelo* o una *trampa iónica* para formar un QTOF o QTRAP respectivamente.

   - Los **Analizadores de sector** (magnético o eléctrico) aceleran los iones de analito que al atravesar el sector son sometidos a un campo

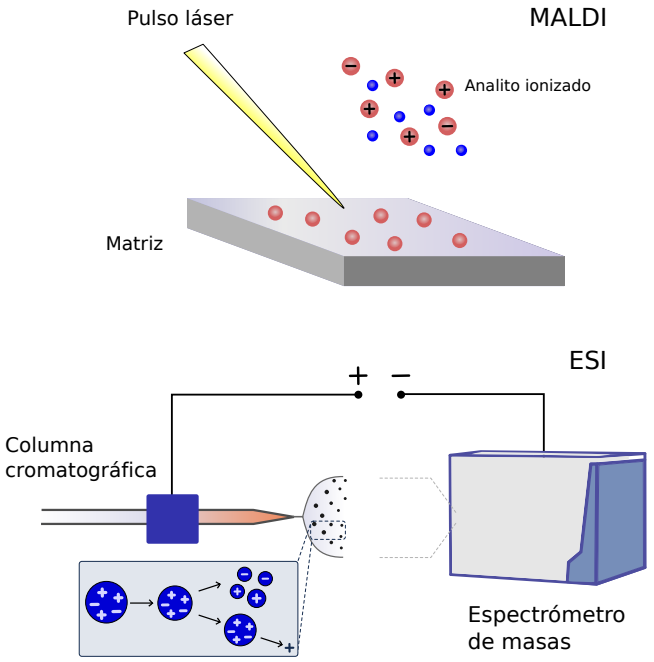


Figura 3: Ionización MALDI y ESI

con fuerza ortogonal a la trayectoria del ion lo que provoca que se desvíen en función de su relación $^{m}/_{z}$

- **Analizadores TOF**, *Time Of Flight*, Tiempo de Vuelo. 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 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. Para iones con la mis-

ma carga, la aceleración depende solo de la masa, los más ligeros llegan al detector antes y los más pesados después. 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.

- **Cuadrupolos**. Los analizadores de tipo Cuadrupolo 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. Una conformación frecuente cosnsite utilizar tres cuadrupolos consecutivos. *bla bla etapas de filtrado, mrm, bla bla*

- Las **Trampas Iónicas** funcionan bajo el mismo principio físico que los cuadrupolos, 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** 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, LIT o LTQ** consisten en un sistema de cuadrupolo, que sitúa los iones en un eje radial, y dos

electrodos terminales, uno en cada extremo, que confina los iones longitudinalmente. Con respecto a las trampas iónicas tipo QIT, las trampas iónicas lineales LTQ tienen una mayor capacidad de acumulación de iones y de barrido. Además pueden funcionar como un cuadrupolo

Orbitrap es un tipo de trampa iónica, relativamente reciente, desarrollado a finales de los años 90 del s.XX (Referencia Makarov) 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.

FTICR

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 mas 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. En principio, pueden utilizarse tantas placas como se quiera, aunque generalmente se utilizan entre 10 y 16. Por medio de este detector, se consiguen amplificaciones dela corriente iónica con factores de multiplicación de 10^6 o mayores

La *sensibilidad*, *resolución* y *precisión* son parámetros importantes en un espectrómetro de masas ya que determinan notablemente la cantidad y calidad

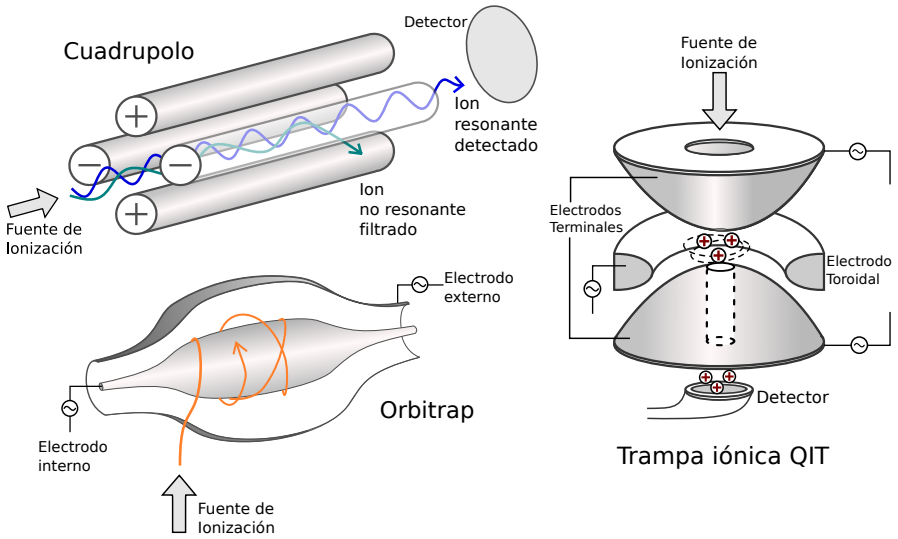


Figura 4: Cuadrupolo, Trampa Iónica Tridimensional(QIT) y Orbitrap

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 millon (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.

La *exactitud* o *precisión* de la medida de la masa molecular es la diferencia entre la masa medida y la calculada, teórica.

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

En la espectrometría de masas en tandem (abreviada MS/MS o MS^2), los péptidos, una vez ionizados y dentro del espectrómetro, son sometidos a una fragmentación adicional. (Figura **??**) 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 MS^1, valores $^{m}/_{z}$ y sus correspondientes intensidades para cada ion. A continuación, en función de la

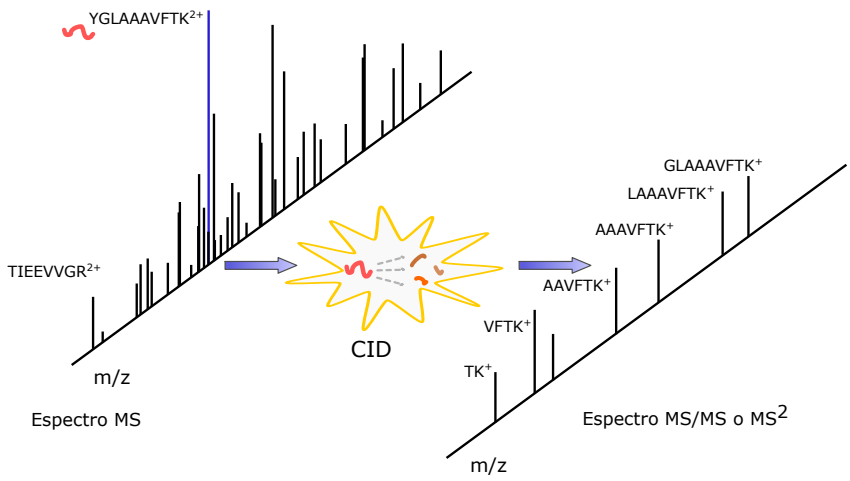


Figura 5: Espectrometría de masas en Tandem. MS/MS

intensidad registrada en MS^1, se seleccionan y aislan algunos de estos iones -*precursores*- para ser fragmentados en péptidos más pequeños -*fragmentos*- en la cámara de colisión del espectrómetro. El espectro MS^2 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 MS^2 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 MS^n, 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 el péptido precursor es seleccionado para ser fragmentado en base a la intensidad

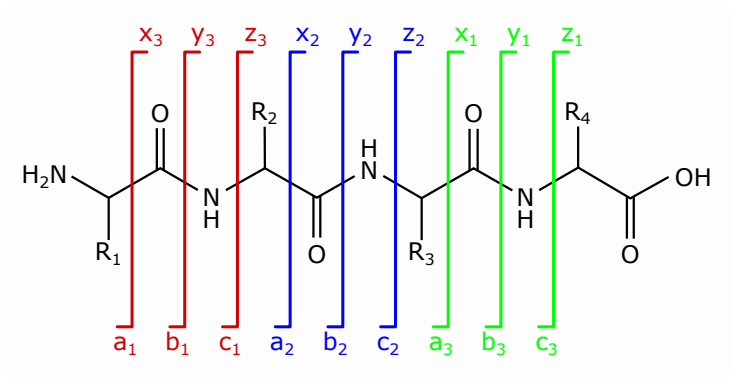


Figura 6: Nomenclatura de Roepstorff para los fragmentos en MS/MS

en el espectro MS[1] se denomina por eso Adquisición Dependiente de Datos (*Data Dependent Acquisition, DDA*)

Los fragmentos originados a partir del péptido, según la nomenclatura de Roepstorff (**?**), 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 6) 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 cuadrupolo-TOF predominan los iones y-, mientras que en las trampas iónicas se generan igualmente b- e y- (**?**)

Técnicas de fragmentación

La *Disociación Inducida por Colisión* (CID) 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 fragmentación tipo CID genera una elevada proporción de iones de las series *b*- e *y*-

En la *Disociación por Transferencia Electrónica* (ETD), (**?**), 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*-. Esta técnica funciona bien para iones de analito con carga $z > 2$ y para péptidos largos e incluso para proteínas enteras. Además produce una fragmentación en la que las cadenas laterales y modificaciones post-traduccionales quedan intactas. Estas cualidades hacen ETD interesante para proteómica *top-down*, que intenta identificar proteínas intactas, y también para experimentos orientados a la deteccion de péptidos con modificaciones post-traduccionales.

En la *Disociación por Captura Electrónica* (ECD) (**?**), como en ETD, la fragmentación se produce por la interacción del analito con electrones que, en este caso, son suministrados por introducción directa. ECD se usa generalmente en espectrómetros tipo FTICR y en trampas iónicas y genera abundantes fragmentos de series *c*-, *z*- y, aunque en menor medida, también de la serie *b*-. Al igual que ECD, es particularmente efectivo para el estudio de péptidos con modificaciones post-traduccionales.

Otro tipo de fragmentación es la *Disociación por Alta Energía de Colisión* (HCD) usado en analizadores tipo Orbitrap (**?**). 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. Este tipo de espectros de fragmentación son más informativos, pero a cambio, requieren analizadores de alta precisión.

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 (**?**), 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. Además la sensibilidad de los espectrómetros es menor para proteínas intactas que para péptidos. La cantidad de posibles formas en que una proteína es procesada, incluyendo modificaciones post-traduccionales en sus péptidos, y variaciones conformacionales entre otras, hace que la combinación de isoformas posibles y sus masas sean imposibles de discernir por MS. Por otra parte, para identificar a la proteína originaria se requiere información de la secuencia y para esto los espectrómetros son mas eficientes si se analizan secuencias de un tamaño limitado en número de aminoácidos.

A pesar de estas limitaciones los espectrómetros sí permiten inferir, al menos parcialmente, secuencias a partir de proteínas intactas y con ello identificarlas. Este es el objetivo de la llamada Proteómica *top-down* (de arriba a abajo)

Sin embargo la proteómica *bottom-up* (*de abajo a arriba*), en la que se infiere la presencia de proteínas a partir de sus péptidos, es la técnica más extendida.

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.

Proteómica en gel

La separación de proteínas en geles de poli-acrilamida (PAGE, *Polyacrilamide Gel Electrophoresis*), 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

- Geles monodimensionales, 1D-PAGE. En este tipo de geles las proteínas se separan en función de su peso molecular. La electroforesis hace que las proteínas mas pequeñas, de menor peso molecular, se desplacen mas lejos en el gel, sometido a una diferencia de potencial.

- Geles bidimensionales. 2D-PAGE. En este caso, las proteínas se separan en una primera dimensión en función de su punto isoeléctrico. Las proteínas se

desplazan sobre una tira con un gradiente de pH hasta situarse en un punto donde su carga neta se equilibra con la de su entorno. A continuación la tira se coloca en la cabecera de un gel y se aplica la segunda dimensión, de modo que se las proteínas se separan más, en este caso por peso molecular, al igual que en un gel 1D.

Otra clasificación posible de las técnicas PAGE puede establecerse en función de si se usan condiciones desnaturalizantes o no

- Geles desnaturalizantes. SDS-PAGE
- Condiciones nativas o no desnaturalizantes Blue Native

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 (**?**). 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 modificaciones post-traduccionales, lo que difícilmente se puede conseguir con otro tipo de aproximaciones.

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 identificarla. 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 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 de Tiempo de Vuelo (TOF).

Una vez obtenido el espectro patrón de masas peptídicas, el proceso de análisis consiguiente es similar al que se hace en la proteómica de alto rendimiento o *shotgun*. Como se describe a continuación, 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 trípticos teóricos.

Proteómica de alto rendimiento *Shotgun*

La llamada proteómica *shotgun* es la técnica de elección para la mayoría de estudios proteómicos a gran escala. 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 a continuación.

Separación multidimensional de péptidos

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

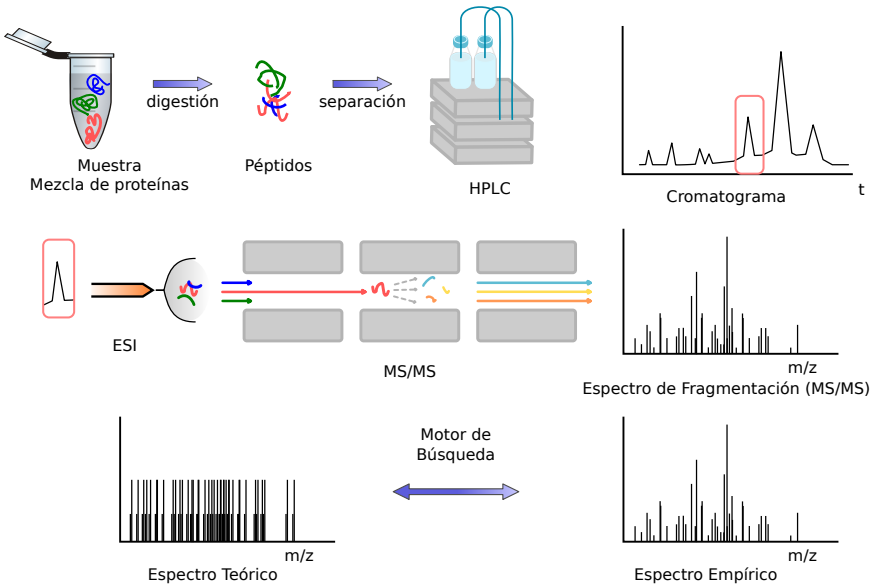


Figura 7: Etapas en un experimento de Proteómica *Shotgun*

experimento, se parte de una muestra más compleja. Esto es importante porque, contando con 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 por ejemplo, por fraccionamientos sub-celulares correspondientes a distintos orgánulos.

Pero el fraccionamiento más importante se hace a nivel de péptido, tras la digestión de las proteinas, por Cromatografía Líquida de Alto Rendimiento (HPLC, *High Performance Liquid Chromatography*). El funcionamiento básico general en HPLC consiste 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 *en Fase Reversa* (RP-HPLC, *Reverse Phase HPLC*) 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 (inversdamente proporcional a la proporción de fase acuosa) provoca que los analitos mas polares eluyan primero integrados cuando la proporción de fase acuosa es más elevada mientras que los más hidrofóbicos son retenidos durante más tiempo.

A continuación, tras la adquisición experimental de los espectros, el paso siguiente en un experimento de proteómica *shotgun* implica el análisis computacio-

nal de esos espectros cuyo objetivo final es la obtención de una lista de proteínas que presumiblemente 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, la inferencia de las proteínas a partir de esos péptidos y una evaluación estadística que aporta medidas de fiablidad a la identificación.

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 asignaciones compuestas por un par péptido-espectro se les denomina generalmente PSM (*Peptide Spectrum Match*)

Las estrategias utilizadas para la asignación de PSMs 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 nuevos espectros adquiridos en el experimento.

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. Existen una gran variedad de herramientas computacionales para realizar esta tarea.

Los motores de búsqueda, descritos en detalle en una sección posterior, son un tipo de programas informáticos a los que 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 se obtienen mediante la predicción de los valores $^{m}/_{z}$ de los péptidos y fragmentos que se generan teóricamente conociendo el patrón de corte de la enzima proteolítica utilizada, las masas de cada aminoácido, 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 han generado el espectro MS/MS y a continuación se 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:

- Masa mono-isotópica o Masa promedio. A partir de un espectro MS se obtiene el valor $^{m}/_{z}$ y con ello la masa del péptido. Sin embargo este valor de masa del péptido puede aproximarse más a la masa mono-isotópica, aquella en la que se considera que todos los átomos de C se encuentran en su forma ^{12}C o bien, puede considerarse que existe una proporción variable de isótopos ^{13}C, y calcularse una masa promedio. Generalmente para espectros de alta resolución la masa del péptido calculada suele acercarse más al valor mono-isotópico, mientras que para instrumentos de baja resolución suele elegirse el valor de masa promedio.

- El número de puntos de corte no efectuado permitidos dentro de la secuencia del precursor. La eficiencia de las enzimas proteolíticas no es del 100 %, por tanto, este parámetro hace que el motor de búsqueda tenga también en cuenta péptidos con K y/o R en su secuencia.
- Modificaciones post-traduccionales y otras modificaciones permitidas que ocurren en el proceso experimental. Algunas de éstas ocurren con una frecuencia variable, como las oxidaciones que se pueden producir en algunas de las metioninas (M). Otro tipo de modificaciones son las llamadas fijas, como la carbamidometilación en las cisteínas (C), un artefacto que se produce en todas las C para evitar la formación de puentes di-sulfuro. La selección de este tipo de modificaciones permite que el motor de búsqueda pueda tener en cuenta un posible desfase de masa en el péptido, equivalente por ejemplo, a un átomo de oxígeno o un grupo carbamido-metil en el caso de las modificaciones descritas.
- Tipo de iones fragmento permitidos. Los espectros de fragmentación teóricos son calculados a partir de las secuencias de péptidos (trípticos) en la base de datos de referencia. Pero los tipos de fragmentos en un espectro MS/MS no son igualmente abundantes sino que, en función del tipo de instrumento y de fragmentación, se generan más fragmentos de un tipo o de otro. Así, en instrumentos de tipo cuadrupolo o híbridos cuadrupolo-TOF, la fragmentación genera abundantes iones de la serie *y*-, en las trampas iónicas además se producen igualmente abundantes iones *b*-.

El establecimiento de estos valores tiene consecuencias muy notables en los resultados de identificación de péptidos y en consecuencia de proteínas. Por ejemplo, restingir 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 refi-

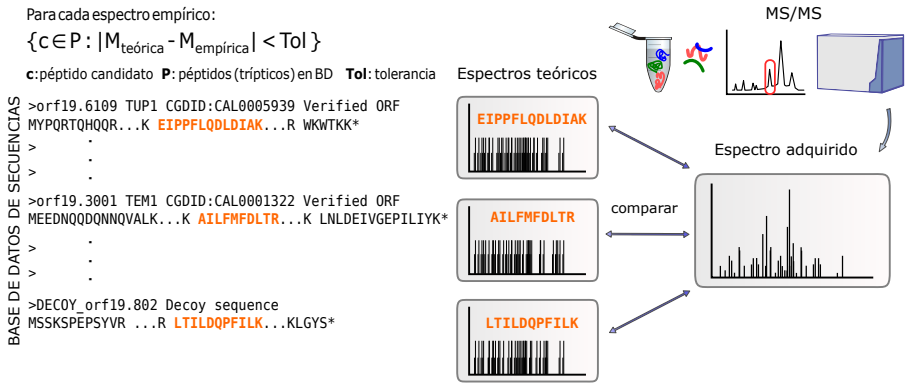


Figura 8: Estrategia básica de identificación. Selección de péptidos candidatos. Correlación espectro MS/MS - secuencia aminoacídica

narla.

Motores de búsqueda. Funciones de puntuación

Los motores de búsqueda se encargan de asignar a cada espectro 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 a nivel de péptido y de fragmento, 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. A continuación se establece el grado de similitud de cada espectro con los espectros teóricos de cada uno de 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. Existen 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 (?) 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 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 elimnados 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 funcion 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 $^{m}/_{z}$) 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.

Otro motor de búsqueda frecuentemente utilizado es *X!Tandem* (?), 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 se tiene en cuenta si los iones coincidentes pertenecen a las series *b*- e *y*-.

Mascot es quizá el más popular de los motores de búsqueda a pesar 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

COMET ?! Ya que se usa en PeptideAtlas2... nombrarlo al menos??

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 (Referencia), Bibliospec( )

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 (Referencia) 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*, 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, las 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.

Búsqueda mediante etiquetas de secuencia

Mirar Hybrid approaches (3.4) en Nesvizhskii review, también -Computational and Statistical Analysis of Protein Mass Spectrometry Data- de MacCoss

Tag-based methods occupy an appealing middle ground between database search and de novo methods. Here, the basic idea is to use de novo analysis to identify a collection of subpeptides (tags) that are hypothesized to occur in the sequence, and then extract candidates from a database that contain the tags. Tag-based methods can be quite fast, and retain the ability to partially identify spectra for which the corresponding peptide is not in the database

Nesviz: -Hybrid approaches are particularly useful for the identification of post-translationally or chemically modified peptides-

Búsquedas tolerantes y multi-etapa

Mirar Strategies for more comprehensive... (4) en Nesvizhskii review

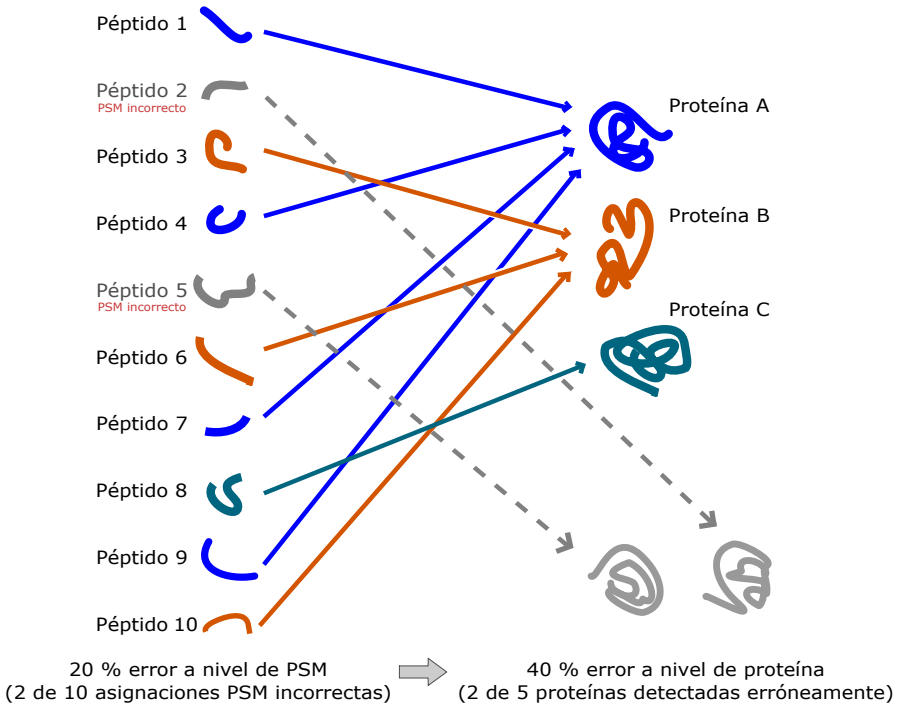


Figura 9: Agrupamiento no aleatorio de péptidos en proteínas

Inferencia de proteínas a partir de péptidos

En un experimento de proteómica *shotgun*, desde el momento de la digestión de las proteínas en péptidos aparece el problema de la trazabilidad en el sentido inverso, es decir inferir la presencia de la proteína originaria de los péptidos que son identificados a partir de los espectros.

Citar aquí PAnalyzer Gorka y cía, los distintos tipos de grupos

Evaluación estadística de los resultados

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 PSMs 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 PSMs incorrectos pueden llegar a suponer la mayoría.

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. Para hacer este tipo de análisis estadístico de los resultados es importante conocer la distribucion de puntuaciones de los péptidos candidatos que pueden asignarse a cada espectro individual y, una vez seleccionado el mejor candidato, la distribucion de todas esas mejores puntuaciones de todos los PSMs en el conjunto del experimento.

Distribuciones de Espectro Individual y Promedio

Distribución de espectro individual.

Cuando un espectro MS/MS es comparado con los espectros teóricos de los péptidos en la lista de candidatos creada a partir de la base de datos de referencia, se puede obtener una lista de péptidos ordenados en función de la puntuación obtenida que mide el grado de similitud entre el espectro empírico adquirido y el teórico. La función de puntuación es un valor dependiente del motor de búsqueda utilizado, por eso, frecuentemente este valor se convierte a una medida estadística más generalizable, el *p-valor* o el *e-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 de todo el resto de péptidos comparados con el

espectro. Esta distribución representa la hipótesis nula, una asignación PSM por azar. 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 por azar 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. Se calcula como el área bajo la curva situada hacia el extremo de la distribución desde el punto de máxima puntuación.

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.

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 PSMs con *p-*

valores igualmente bajos solo por azar. Por este motivo se utilizan estrategias de *corrección de test múltiple* (*multiple test correction*) que re-ajustan los *p-valores*. Una aproximación muy utilizada, aunque produce resultados conservadores, es la corrección de Bonferroni((?), 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 PSMs, el *p-valor* original habría de reajustarse a $0{,}05/10.000 = 5{\cdot}10^{-6}$.

Las distribuciones promedio, como muestra la Figura 11, son distribuciones de las mejores puntuaciones de todos los PSMs 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 Error Posterior, PEP.

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 PSMs que han sido ordenados por *p-valores* o *e-valores* obtenidos individualmente.

Bases de datos señuelo 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)* (?). Básicamente, el concepto de tasa FDR se refiere a la proporción de PSMs incorrectos que se aceptan en todo el conjunto de PSMs de un experimento para un umbral de puntuación (o de parámetro estadístico como el *p-valor*) fijado.

Para la estimación de la tasa FDR (Figura 11), la estrategia utilizada consiste esencialmente en utilizar una base de datos llamada *señuelo* o *decoy* (?). 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

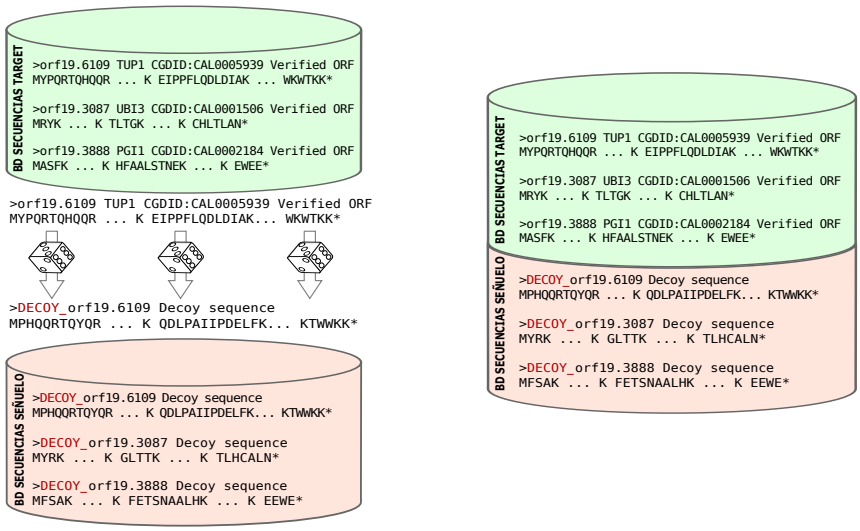


Figura 10: Construcción de una base de datos señuelo

que no existen, no corresponden a ninguna proteína (Figura 10) 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 (**??**) 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 se-

cuencias *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 PSMs 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 PSMs *señuelo* (N_d, espectros a los que se ha asginado una secuencia *señuelo*) Con esto se puede estimar la tasa FDR como N_d/N_t, esto es, la proporción de PSMs *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 *se-*

ñuelo a pesar de que algunos de los espectros podrían asignarse correctamente a una secuencia *target*. Además, la mayoría de las funciones de puntuación tienden a otorgar puntuaciones más altas a PSMs *señuelo* que a PSMs *target* incorrectos por lo que la distribución de puntuaciones *señuelo* no es un reflejo preciso de la distribución de puntuaciones de los PSMs 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 PSMs con puntuaciones bajas son probablemente incorrectos (**?**) Así se puede incluir en la tasa FDR un factor de corrección definido por el porcentaje estimado de PSMs *target* incorrectos (PIT): Si en N_t el 80 % de los PSMs 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 PSMs *señuelo* en el conjunto de PSMs aceptado se estiman 80 PSMs *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 PSMs correcto y por tanto un incremento de FDR

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

IMPORTANTE: Mencionar MAYU, explicar en que consiste un poquito por lo menos (Y quiza tambien Percolator

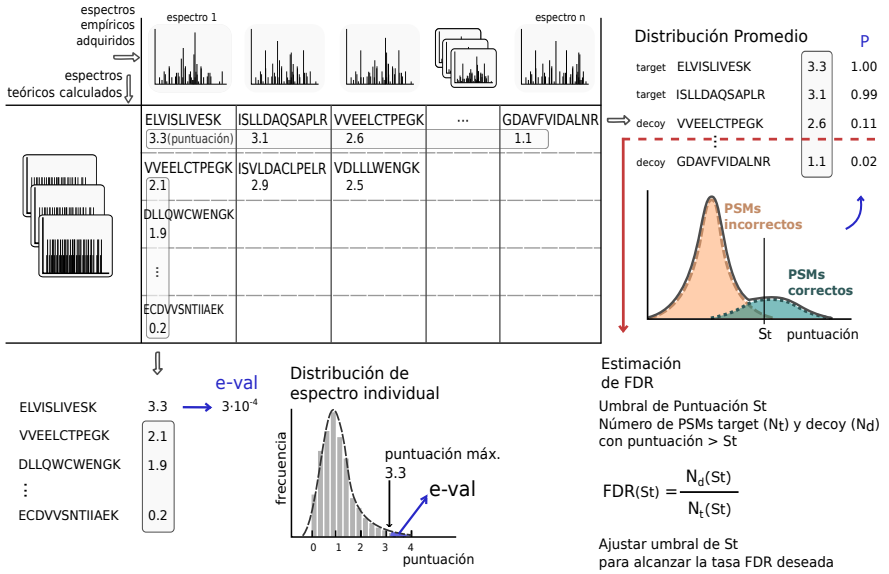


Figura 11: Distribuciones de Espectro Individual y Promedio

Modelos mixtos de probabilidad. Probabilidad de Error Posterior (PEP). PeptideProphet

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. Esto se puede calcular mediante la *Probabilidad de Error Posterior, PEP*, de un PSM, que se define como la fracción estimada de PSMs correctos entre una colección de PSMs con puntuaciones (o p-valores / e-valores) similares. Este tipo de análisis, propuesto por (**?**) e implementado en la herramienta *PeptideProphet*, se basa en asumir que la distribución promedio de puntuaciones de todos los PSMs del experimento es una combinación de distribuciones, un modelo mixto de probabilidad, que integra la distribución de los PSMs incorrectos y la distribución de los PSMs correctos. (Figura 11)

En un principio, este modelo mixto para el cálculo de las PEPs ajustaba un modelo paramétrico a la distribución de puntuaciones de los PSMs de forma no supervisada y no requería secuencias *señuelo* Pero con la incorporación de PSMs *señuelo*

The Probability Ratio Method, Navarro et al

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 experimentos 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 (**?**), 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

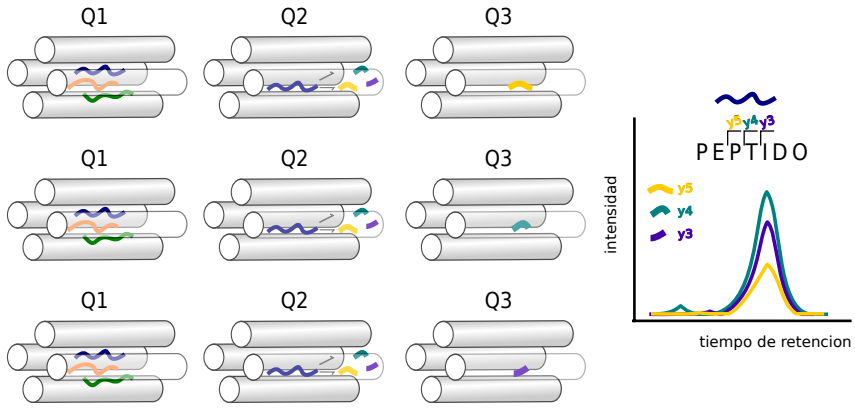


Figura 12: Adquisición y reconstrucción de la señal en un experimento SRM

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

Candida albicans como organismo modelo

Repositorios publicos de proteómica shotgun y dirigida

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 elaboracion de listas de proteínas identificadas 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.

La iniciativa HUPO-PSI tiene un papel muy importante en la elaboración y adopción de formatos abiertos que puedan servir de estándar para toda la comunidad. En esta tarea de estandarización están implicados representantes de las compañías fabricantes de equipos, editores de revistas, y desarrolladores de código del ámbito académico para revisar y proporcionar formatos que recojan toda la información necesaria (MIAPE)

En una clasificacion 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 en que el instrumento registra los espectros.

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 datos perfil permiten ver la forma de los picos del espectro de masas. Cada unidad de masa atómica se divide en muchos intervalos de muestreo. La intensidad de la corriente iónica se determina en cada uno de los intervalos de muestreo. Este tipo de datos muestra la intensidad en cada uno de los intervalos de muestreo con las intensidades conectadas por una línea continua. En otras ocasiones, con el objetivo de ahorrar espacio, solo se registra la señal cuando se supera un cierto umbral de intensidad que discierne la señal real del ruido. Este tipo de espectros es similar a los adquiridos en modo continuo pero los valles entre picos no quedan registrados.

Los instrumentos registran los espectros en modo continuo de forma predeterminada, pero frecuentemente son sometidos a un procesamiento por un algoritmo que extrae los picos detectados como parejas de valores $^{m}/_{z}$ e intensidad. Estos son los datos centroide, muestran el espectro de masas como un gráfico de barras y suman las intensidades de cada conjunto de varios intervalos de muestreo. Esta suma se muestra frente al centrado integral de masa de los intervalos de muestreo.

- *Formatos que recogen el resultado de las búsquedas*
- *Formatos que almacenan bibliotecas de espectros*
- *Formatos que almacenan secuencias*
- *Formatos específicos para proteómica dirigida*

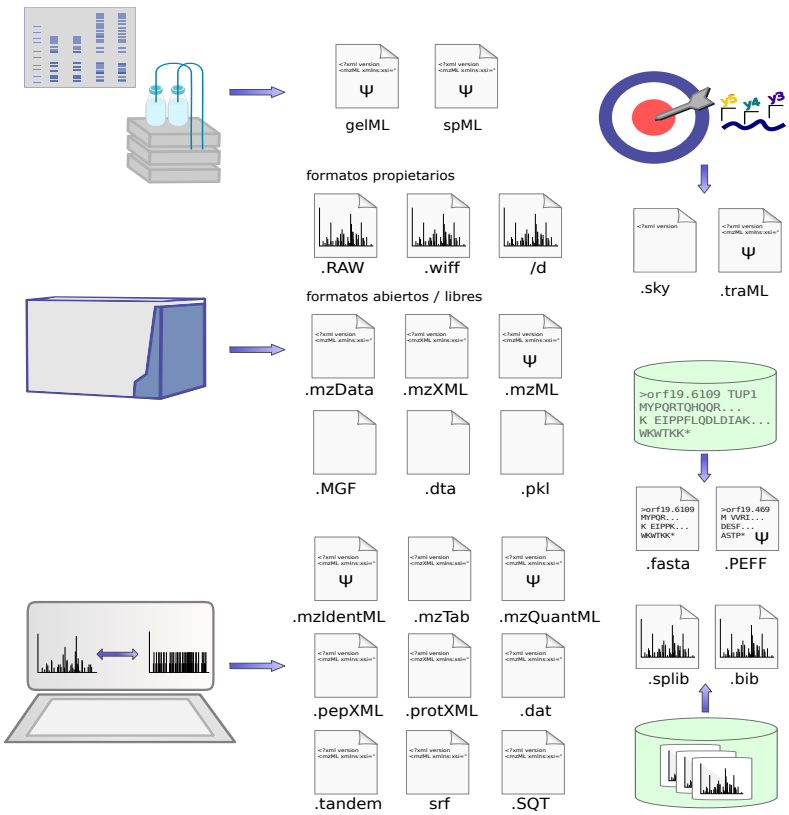


Figura 13: Visión general de formatos comunmente usados en cada etapa de un experimento de proteómica

Desarrollo de una aplicacion web para datos de proteómica shotgun de *Candida albicans*

Capítulo 1

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

 DOI 10.1002/pmic.200900023 *Proteomics* 2009, *9*, 4664–4668

TECHNICAL BRIEF

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

Vital Vialás[1], Rubén Nogales-Cadenas[2], César Nombela[1], Alberto Pascual-Montano[2] 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 overproliferate 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 publish 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 they 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]	*C. albicans* differentially expressed proteins after 3 h interaction with RAW 264.7 murine macrophages. 2-D silver-stained gel. MS/MS (MALDI/TOF-TOF)	*C. albicans*	66
[3]	Proteins identified from cytoplasmic extracts of RAW 264.7 cells after 45 min interaction with *C. albicans*	*Mus. musculus*	38
[13]	Identification of Glycosyl phosphatidil inositol (GPI)-anchored membrane proteins	*C. albicans*	292
	Identification of membrane proteins		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

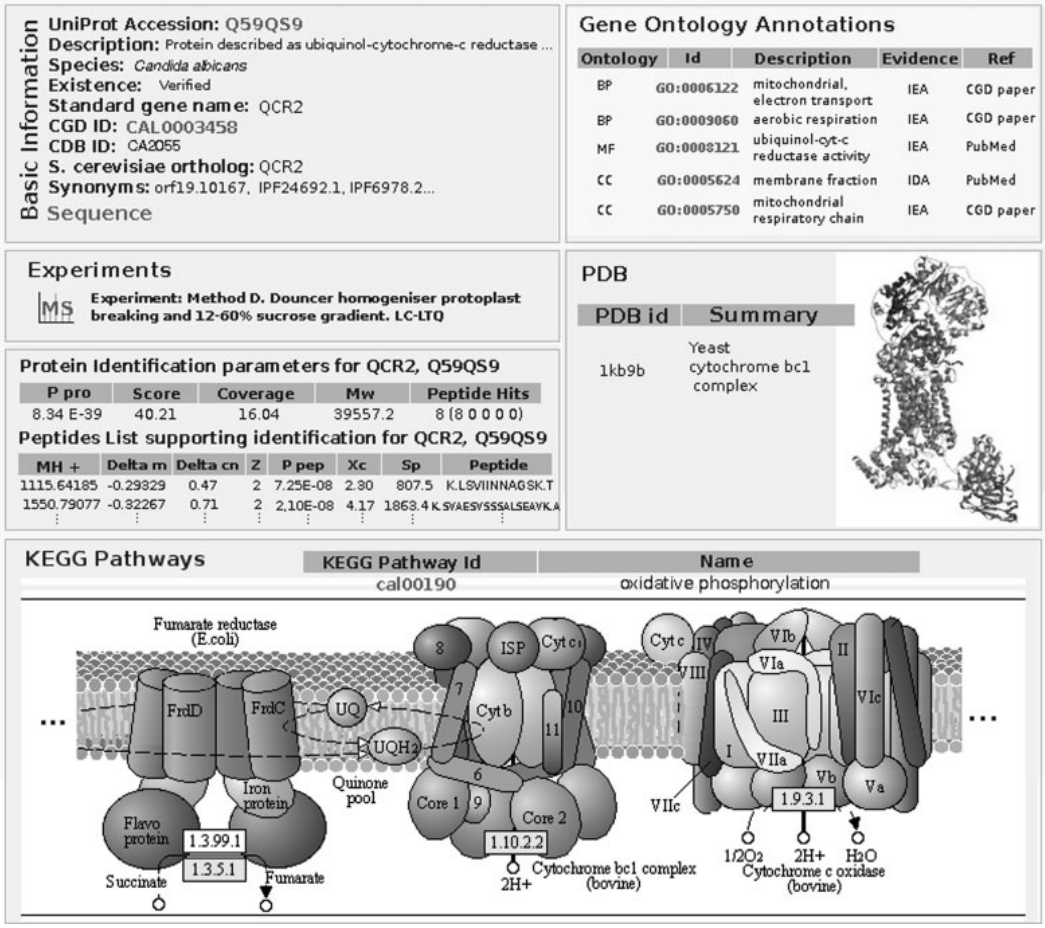


Figure 1. Use case: Search for *C. albicans* ubiquinol-cytochrome-*c* reductase QCR2. The different sections in the result comprise information on protein description and identifiers, experiments in which it has been identified, GO annotation, KEGG and CGD pathway annotation and structural information 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. Douncer 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 Comision Interministerial de Ciencia y Tecnología (CYCIT, Spain), DEREMICROBIANA – 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.

...

...

Capítulo 2

Proteopathogen 2, adaptación al formato estándar de identificaciones .mzIdentML

2.1.

...
...

Creación de un PeptideAtlas de *Candida albicans*

Capítulo 3

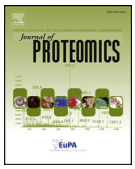
A *Candida albicans* PeptideAtlas

JOURNAL OF PROTEOMICS 97 (2014) 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,*]

[a]*Dept. Microbiología II, Universidad Complutense de Madrid, Madrid, Spain*
[b]*IRYCIS: Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain*
[c]*Institute for Systems Biology, Seattle, WA, USA*
[d]*CSIC/UAB Proteomics Laboratory, Instituto de Investigaciones Biomédicas de Barcelona—Consejo Superior de Investigaciones Científicas, Spain*
[e]*Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland*
[f]*Faculty 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.

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: vvialasf@ucm.es (V. Vialas), conchagil@ucm.es (C. Gil).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.06.020

1. Introduction

Candida albicans is a fungus of great clinical importance. In addition to asymptomatically 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*

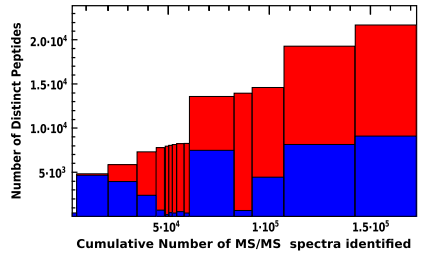


Fig. 1 – Histogram showing the cumulative number of distinct peptides in the *C. albicans* PeptideAtlas. Each bar represents a different experiment that has contributed to the build. Bar width is proportional to the number of high confidence PSMs. Height of the blue section of the bar represents the 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

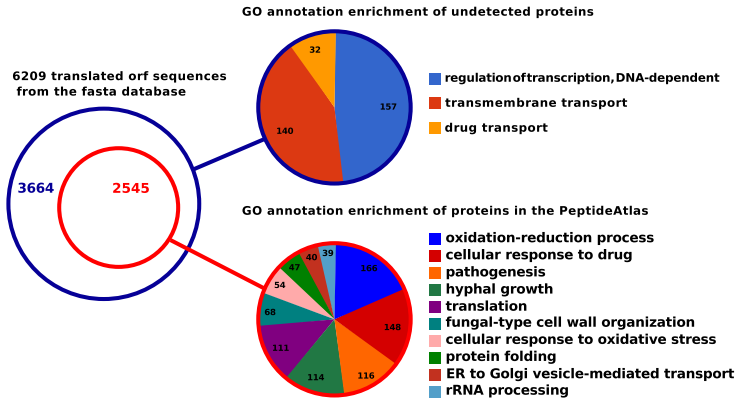


Fig. 2 – Gene Ontology annotation enrichment analysis for both the covered and undetected proteome subsets. All shown GO annotations correspond to the biological process ontology and were found significant for a p-value cut-off below 0.01.

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 Utr2p; 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 Mca1p, 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

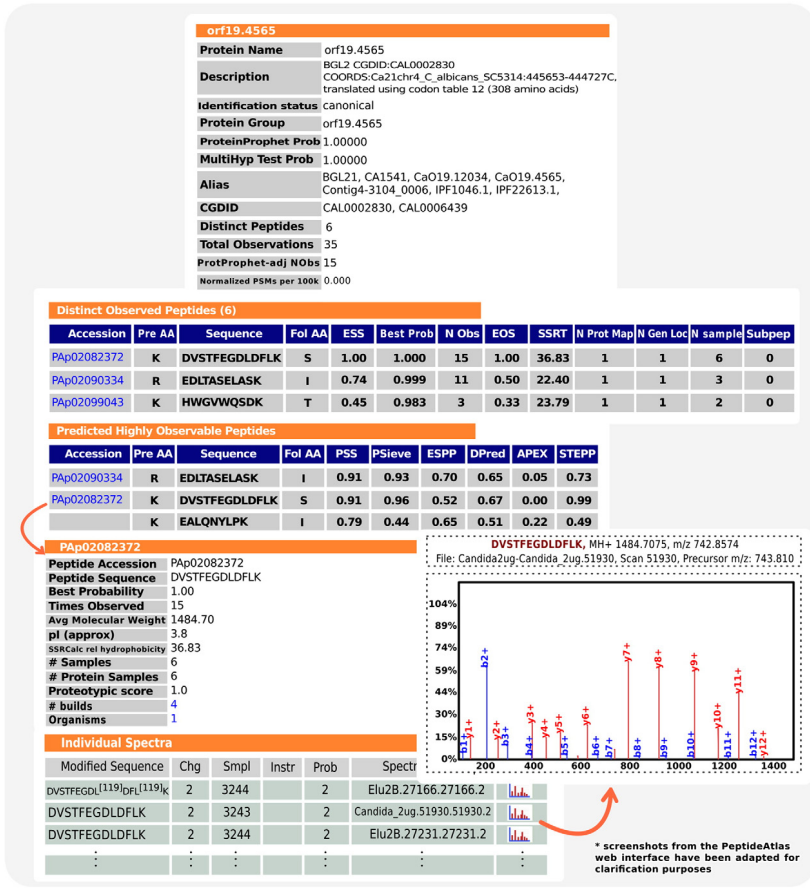


Fig. 3 – Protein- and peptide-centric views for Bgl2p are depicted. *Distinct observed peptides* are ranked by the *BestProb* parameter (representing the PeptideProphet probability). Of those, most probably, some will also be present in the following *Predicted Highly Observable Peptides* table were 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=323and 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 Hernáez 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.

REFERENCES

[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 [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 [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 [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 [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 [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 [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 [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-Embún 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.

3.1.

...

...

Capítulo 4

Incremento en la cobertura del proteoma en el PeptideAtlas de Candida albicans

4.1.

...
...

Desarrollo de una base de datos para datos de Proteómica Dirigida (MRM)

Bibliografía

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

Miguel de Cervantes Saavedra

–¿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 del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes

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

Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes