
Desarrollo de herramientas bioinformáticas para Proteómica de alto rendimiento



TESIS DOCTORAL

Vital Vialás Fernández

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

2015

Desarrollo de herramientas bioinformáticas para Proteómica de alto rendimiento

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

Dirigida por la Doctora
Concha Gil García

Departamento de Microbiología II
Facultad de Farmacia
Universidad Complutense de Madrid

2015

Dedicatoria

Agradecimientos

Frase célebre que vaya bien en los agradecimientos

Gracias y bla bla

Resumen

Introducción

La Proteómica a Gran Escala, también llamada de Alto Rendimiento, basada en la separación de proteínas (mediante electroforesis) y de péptidos (mediante cromatografía) seguida de espectrometría de masas en tandem, es la técnica fundamental de la Proteómica moderna. En los experimentos de Proteómica, a continuación de la parte experimental que conduce a la adquisición de los espectros, los resultados se obtienen mediante la identificación de los péptidos y la inferencia de las proteínas originarias presentes en las muestras. Para ello, los distintos motores de búsqueda emplean diferentes estrategias y a continuación generalmente se evalúan estadísticamente los resultados mediante métodos como la Tasa de Falsos Descubrimientos, FDR o modelos estadísticos más complejos.

Objetivos

La presencia de resultados de experimentos de Proteómica en repositorios públicos para el hongo patógeno oportunista *Candida albicans* eran hasta hace poco muy escasos, originados en instrumentos de baja resolución y por tanto no muy fiables. Así, el desarrollo de bases de datos y adopción de formatos estándar en Proteómica juegan un papel esencial para analizar, comparar y presentar resultados. Las herramientas informáticas descritas en esta tesis contribuyen a esos objetivos.

Resultados

La base de datos y herramienta web Proteopathogen es la primera aplicación online descrita que combinaba resultados de experimentos de Proteómica con información específica relevante para el estudio de proteínas de *C. albicans* como términos de la ontología génica GO (*Gene Ontology*) o las rutas de la Enciclopedia de Genes y Genomas de Kyoto, KEGG (*Kyoto Encyclopedia of Genes and Genomes*) en las que están implicadas. Tras una primera versión en que los resultados de los experimentos se recogían en formatos de texto separado

RESUMEN

por tabulador dependientes de la herramienta de *software* que hubiera generado los resultados (Vialás et al., 2009), la aplicación y la base de datos ha sido totalmente remodelada para adaptarse al formato estándar de identificaciones mzIdentML (Vialas and Gil, 2015) como fuente de información. Proteopathogen es actualmente una herramienta *online* que facilita la visualización y análisis de los resultados de experimentos de Proteómica tanto por los usuarios del laboratorio como por parte de los revisores de las revistas donde los estudios son publicados.

Paralelamente, el desarrollo de un PeptideAtlas para *Candida albicans* ha supuesto la mayor caracterización de su proteoma disponible hasta la fecha describiendo más de 71000 péptidos detectados y asignados a 4174 proteínas lo que supone un 66 % del proteoma predicho. Es además el primer proteoma de un modelo de hongo patógeno en el proyecto global PeptideAtlas y cuenta con la robustez y fiabilidad en las identificaciones que proporciona el flujo de análisis TPP (*Trans Proteomics Pipeline*).

Summary

Introduction

The concept of Proteomics, a term coined in analogy to Genomics, was first used by Marc Wilkins in the mid 90s to describe the total set of proteins being expressed by the genes of a cell, tissue or organism. Earlier, in the late 80s, the development of soft ionizations techniques, such as Electrospray Ionization (ESI) and Soft Laser Desorption (SLD) enabled the ability to ionize large bio-molecules such as proteins while keeping them relatively intact. This settled the foundation upon which Mass Spectrometry was applied to what would later become modern Proteomics.

In *shotgun* Proteomics, the first step of the experiment usually consists of a digestion of the sample proteins into peptides by means of a proteolytic enzyme such as trypsin. This greatly increases performance in terms of the number of proteins that can be detected in a single LC-MS/MS run compared to gel-based approaches, but comes at the cost of a great complexity at the peptide level and the protein inference problem.

Peptides are then separated by liquid chromatography (LC), then ionized and eventually enter the mass spectrometer where they are separated as a function of their mass-to-charge (m/z) ratio, recorded in the MS1 spectrum. In tandem mass spectrometry (MS/MS) peptides with higher intensities are selected to be fragmented so MS/MS spectra, a collection of m/z values and intensities of the precursor and product ions, are produced.

Once the empirical spectra are acquired, the computational analysis starts. The most efficient peptide identification method is based on searching the acquired spectra against protein sequence databases. This is what search engines do. Basically, a score measuring the degree of similarity between the empirical spectrum and a theoretically derived spectrum (corresponding to a known sequence) is given to pairs of spectrum - peptide sequence named PSMs (Peptide to Spectrum Match)

To assess confidence in peptide identification statistic measures such as p-values and e-values are given to PSMs. But in the context of a large experiment generating thousands of MS/MS spectra, further filtering or additional statistical procedures may be applied.

SUMMARY

The False Discovery Rate Probability PeptideProphet

Then there is the protein inference problem: 2 problems addressed by ProteinProphet non-random peptide-to-protein grouping which explodes FDR levels and peptide degeneracy, that is, conserved sequences in different proteins

Objectives

The presence of proteomic data related to the fungal opportunistic pathogen *Candida albicans* on online public repositories was until recently very sparse, sometimes originated from low resolution instruments and therefore not very robust. In this context, the development databases and the adoption of standard formats in Proteomics have an essential role in the analysis, sharing and dissemination of results. The software tools presented here contribute to these objectives.

Results

The database and web tool Proteopathogen was the first described software that combined Proteomics experiments results with specific information relevant to the study of *C.albicans* proteins such as GO (Gene Ontology) terms and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways annotations. Following a first version where the results were collected in a tab separated text format dependant on the software used to generate them (Vialás et al., 2009), the software was adapted to make use of the standard format for protein and peptide identifications, mzIdentML, as its source of data (Vialas and Gil, 2015). Proteopathogen is up to date an online tool to allow visualization and analysis of Proteomics experiments results by the wet-lab users but also enabling revision by reviewers of the journals of the field.

At the same time, the development of a *Candida albicans* PeptideAtlas means the most exhaustive proteome characterization up to the current date. It describes over 71000 detected peptides assigned to 4174 proteins, which represent 66 % of the predicted proteome. Is is, in addition, the first model of a fungal pathogen present in the global PeptideAtlas project and benefits from the robustness and confidence that TPP (Trans Proteomics Pipeline) enables.

Índice

| | |
|---|------------|
| Agradecimientos | VII |
| Resumen | IX |
| Summary | XI |
| | |
| I INTRODUCCIÓN | 1 |
| 1. Introducción | 3 |
| 1.1. Proteómica. Conceptos generales | 4 |
| 1.2. Espectrometría de masas | 4 |
| 1.2.1. Consideraciones sobre unidades empleadas en espectrometría de masas | 5 |
| 1.2.2. Componentes de un espectrómetro de masas | 6 |
| 1.2.3. Espectrometría de masas en Tándem. MS/MS | 12 |
| 1.3. Digestión de proteínas en péptidos | 14 |
| 1.4. Proteómica en gel | 15 |
| 1.4.1. Huella Peptídica | 16 |
| 1.5. Proteómica a gran escala o de alto rendimiento | 17 |
| 1.5.1. Separación de péptidos y proteínas sin gel | 17 |
| 1.6. Asignación Péptido-Espectro | 19 |
| 1.6.1. Búsqueda en bases de datos de secuencias | 19 |
| 1.6.2. Otras estrategias de asignación péptido-espectro | 24 |
| 1.7. Evaluación estadística de las identificaciones de péptidos y proteínas | 25 |
| 1.7.1. Conceptos estadísticos básicos | 25 |
| 1.7.2. Puntuaciones basadas en distribuciones de espectro individual y promedio | 26 |
| 1.7.3. Bases de datos señuelo y Tasa de Falsos Descubrimientos (FDR) | 30 |
| 1.7.4. Modelos mixtos de probabilidad. Probabilidad Posterior | 32 |

ÍNDICE**ÍNDICE**

| | |
|--|-----------|
| 1.8. Inferencia de proteínas a partir de péptidos | 35 |
| 1.9. Herramientas adicionales de post-procesamiento y validación a nivel de péptido y proteína | 37 |
| 1.10. Proteómica dirigida. SRM/MRM | 39 |
| 1.11. Repositorios públicos de Proteómica online | 39 |
| 1.11.1. PRIDE | 40 |
| 1.11.2. PeptideAtlas | 40 |
| 1.12. Formatos de archivos usados en espectrometría de masas y Proteómica | 41 |
| 1.13. <i>Candida albicans</i> como organismo modelo | 44 |
| | |
| OBJETIVOS | 47 |
| Objetivos | 49 |
| | |
| II DESARROLLO DE UNA APLICACIÓN WEB PARA RESULTADOS DE EXPERIMENTOS DE PROTEÓMICA DE <i>Candida albicans</i> | 51 |
| Proteopathogen, a protein database for studying <i>Candida albicans</i> - host interaction | 53 |
| Proteopathogen2: A database and web tool to store and display proteomics identification results in the mzIdentML standard | 63 |
| Introduction | 66 |
| Materials and methods | 67 |
| Results and discussion | 68 |
| Conclusions | 71 |
| | |
| III CREACIÓN DE UN PEPTIDEATLAS DE <i>Candida albicans</i> | 75 |
| A <i>Candida albicans</i> PeptideAtlas | 77 |
| Introduction | 80 |
| Materials and methods | 81 |
| Results and discussion | 83 |
| Conclusions | 88 |
| | |
| A comprehensive <i>Candida albicans</i> PeptideAtlas build enables deep proteome coverage | 93 |
| Introduction | 96 |
| Experimental procedures | 97 |

ÍNDICE**ÍNDICE**

| | |
|--|------------|
| Results and discussion | 102 |
| Conclusions | 110 |
| IV DISCUSIÓN | 115 |
| Discusión | 117 |
| Desarrollo de una aplicación web para datos de proteómica a gran escala de <i>Candida albicans</i> | 118 |
| Desarrollo de un PeptideAtlas Candida albicans | 120 |
| V CONCLUSIONES | 123 |
| Conclusiones | 125 |
| Bibliografía | 127 |
| Índice alfabético | 141 |
| Lista de acrónimos | 141 |

Índice de figuras

| | |
|--|-----------|
| Introducción | 3 |
| 1.1. Aumento de complejidad desde el genoma hacia el proteoma | 5 |
| 1.2. Esquema de un espectrómetro de masas | 7 |
| 1.3. Ionización MALDI y ESI | 8 |
| 1.4. Analizadores de masas | 10 |
| 1.5. Espectrometría de masas en Tándem. MS/MS | 12 |
| 1.6. Nomenclatura de Roepstorff para los fragmentos en MS/MS | 14 |
| 1.7. Etapas en un experimento de Proteómica <i>Shotgun</i> | 17 |
| 1.8. Interpretación automática de espectros MS/MS | 20 |
| 1.9. Estrategia básica de identificación | 22 |
| 1.10. Tabla de contingencia. Contraste de hipótesis | 27 |
| 1.11. Estimación del e-valor | 28 |
| 1.12. Distribuciones de Espectro Individual y Promedio | 29 |
| 1.13. Construcción de una base de datos sueño | 31 |
| 1.14. Modelo mixto de probabilidad. PeptideProphet | 34 |
| 1.15. Agrupamiento no aleatorio de péptidos en proteínas | 36 |
| 1.16. Adquisición y reconstrucción de la señal en un experimento SRM | 40 |
| 1.17. Formatos de archivos en Proteómica | 42 |
| 1.18. Equilibrio entre <i>Candida albicans</i> y células del sistema inmune | 45 |
| Proteopathogen, a protein database for studying <i>Candida albicans</i> - host interaction | 53 |
| Figure 1 | 57 |
| Proteopathogen2: A database and web tool to store and display proteomics identification results in the mzIdentML standard | 63 |
| Figure 1 | 68 |

ÍNDICE DE FIGURAS

ÍNDICE DE FIGURAS

| | |
|--|-----|
| Figure 2 | 70 |
| A <i>Candida albicans</i> PeptideAtlas | |
| Figure 1 | 83 |
| Figure 2 | 85 |
| Figure 3 | 87 |
| A comprehensive <i>Candida albicans</i> PeptideAtlas build enables deep proteome coverage | |
| Figure 1 | 97 |
| Figure 2 | 103 |
| Figure 3 | 103 |
| Figure 4 | 106 |
| Figure 5 | 110 |

Índice de Tablas

| | |
|--|-----------|
| Proteopathogen, a protein database for studying <i>Candida albicans</i> - host interaction | 53 |
| Table 1 | 57 |
| Proteopathogen2: A database and web tool to store and display proteomics identification results in the mzIdentML standard | 63 |
| Table 1 | 67 |
| A <i>Candida albicans</i> PeptideAtlas | 77 |
| Table 1 | 81 |
| A comprehensive <i>Candida albicans</i> PeptideAtlas build enables deep proteome coverage | 93 |
| Table 1 | 96 |
| Table 2 | 107 |

INTRODUCCIÓN

Introducción

INTRODUCCIÓN

1.1. Proteómica. Conceptos generales

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

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

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

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

1.2. Espectrometría de masas

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

Como ocurre en muchas otras ocasiones en la ciencia, de forma paralela e independientemente habían surgido en distintas partes del mundo ideas muy similares. Así, el desarrollo de

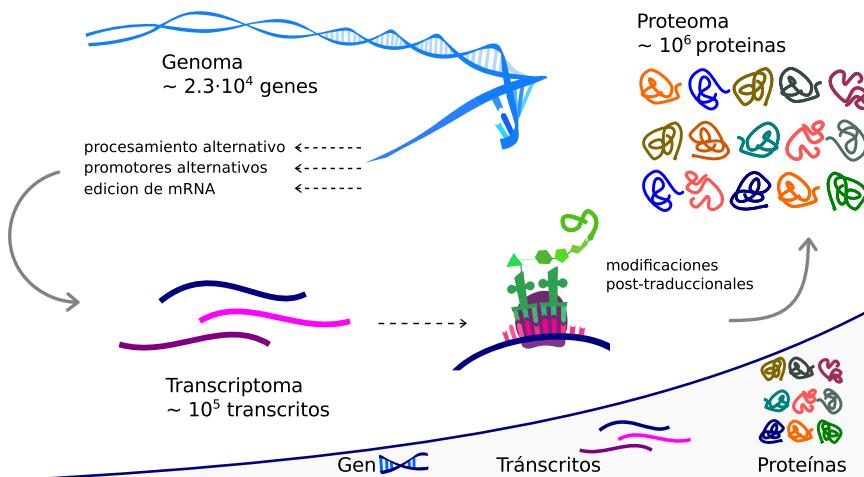


Figura 1.1: Aumento de complejidad desde el genoma hacia el proteoma

SLD que valió el Nobel a K. Tanaka, tuvo un precedente unos años antes. Franz Hillenkamp y Michael Karas en Frankfurt, Alemania (éstos discutiblemente no galardonados) habían ideado una técnica similar que, en este caso, denominaron Ionización mediante Desorción por Láser Asistida por Matriz, MALDI (*Matrix Assisted Laser Desorption Ionization*) (Karas and Hillenkamp, 1988). Aunque MALDI no fue aplicada a la ionización de proteínas hasta la publicación del trabajo de Tanaka, actualmente éste es el acrónimo que se ha impuesto para referirse a la técnica y es, de hecho, una técnica muy extendida en laboratorios de espectrometría de masas.

1.2.1. Consideraciones sobre unidades empleadas en espectrometría de masas

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

INTRODUCCIÓN

un átomo de hidrógeno o la masa de un protón.

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

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

$$1Da = 1u = 1,66054 \times 10^{-24} g/\text{átomo} \ ^{12}\text{C} = 1,66054 \times 10^{-27} kg/\text{átomo} \ ^{12}\text{C}$$

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

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

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

1.2.2. Componentes de un espectrómetro de masas

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

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

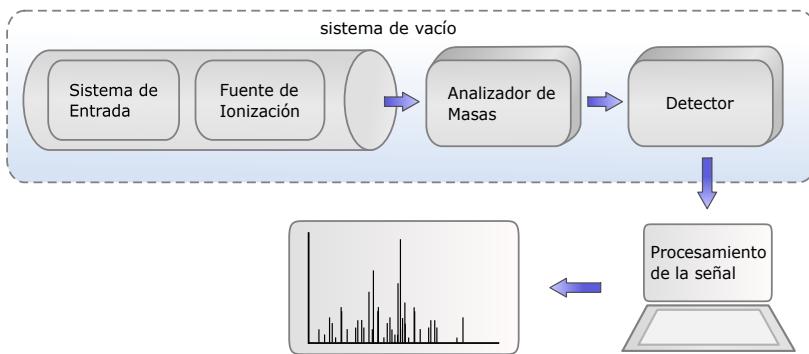


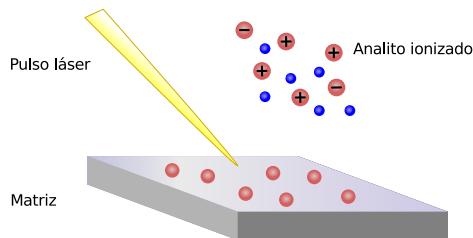
Figura 1.2: Esquema de un espectrómetro de masas. En ocasiones el caso de ESI, el sistema de entrada y la fuente de ionización forman parte de un único componente. Todos los componentes se encuentran en el interior de un sistema de vacío

2. Fuente de iones. Las macromoléculas biológicas, como proteínas y péptidos, no son fácilmente volatilizadas. El desarrollo de las técnicas de ionización suave permitió que péptidos y proteínas ionizados y relativamente intactos pudieran ser introducidos, en fase gaseosa, en un sistema de vacío en los espectrómetros de masas para ser analizados. La ionización ESI y MALDI son las más comunes en Proteómica aunque existen también otros métodos un poco menos utilizados.

- **MALDI** (Figura 1.3 a) consiste en embeber la muestra en una matriz líquida, que posteriormente se seca, con alta capacidad de absorber luz UV sobre la que inciden pulsos de luz láser UV. Al absorber la energía del láser las moléculas que conforman la matriz son ionizadas por adición de protones que son luego transferidos al analito.
- En **ESI** (Figura 1.3 b) el analito se encuentra en fase líquida en un solvente orgánico volátil como metanol o acetonitrilo. Esta solución es conducida a través de un capilar sometido a un campo eléctrico de forma que las micro-gotas en el ápice del capilar, una vez que la carga supera un límite, forman un aerosol. Las micro-gotas del aerosol disminuyen su tamaño por evaporación del solvente, reagrupándose en gotas más estables y pequeñas en un proceso reiterativo, hasta el punto en que las moléculas de analito se repelen con la fuerza suficiente para superar la tensión superficial y liberarse (explosión de Coulomb) quedando en suspensión y siendo así introducidos en un sistema vacío hacia el espectróme-

INTRODUCCIÓN

a) MALDI



b) ESI

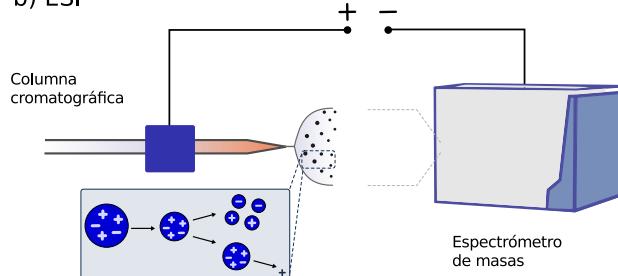


Figura 1.3: Ionización MALDI y ESI

tro.

Otros tipos de ionización menos comunes son el Bombardeo Rápido Atómico, FAB (*Fast Atom Bombardment*), la Desorción por Campo Eléctrico, FD (*Field Desorption*), y la Desorción por Plasma, PD (*Plasma Desorption*).

3. **Analizador de masas.** El analizador de masas es la parte del instrumento en la que los iones se separan en base su relación entre la masa y carga (m/z). Es el elemento que se usa generalmente para definir el tipo de instrumento. Existen varios tipos, y además pueden combinarse en los llamados espectrómetros híbridos. Así, un analizador de tipo cuadrupolo puede encontrarse acoplado con un analizador de *tiempo de vuelo* o una *trampa iónica* para formar un Cuadrupolo-Tiempo de Vuelo, QTOF (*Quadrupole-Time Of Flight*) o Cuadrupolo-Trampa Iónica, QTrap (*Quadrupole-Ion Trap*) respectivamente. Algunos de los analizadores de masas más comunes son los siguientes:

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

a la trayectoria del ion lo que provoca que se desvíen en función de su relación m/z .

- **Analizadores de Tiempo de Vuelo.** TOF (*Time Of Flight*) Este tipo de analizador usa un campo eléctrico para acelerar los iones de analito. La separación se produce por la diferencia en el tiempo que éstos invierten en recorrer una distancia en el vacío en el interior del analizador, el llamado *Tiempo de Vuelo*. La aceleración y por tanto el Tiempo de Vuelo es una función de la relación m/z de los iones que impactan en el detector a diferentes tiempos. Por su carácter dependiente de la dimensión tiempo, los analizadores TOF se usan generalmente en combinación con ionización MALDI, que introduce iones en el analizador en pulsos de láser.
- **Cuadropolos.** Los analizadores de tipo Cuadropolo (Figura 1.4 a) reciben su nombre porque constan de cuatro varillas metálicas enfrentadas en pares llamados polos con cargas opuestas. Sobre estos pares, además del potencial eléctrico de corriente continua, se aplica también una corriente alterna de radiofrecuencia. Esta conformación permite crear un campo eléctrico oscilante controlado que estabiliza (o desestabiliza) selectivamente los iones que pasan a través y de esta forma solo los iones con ciertos valores m/z podrán llegar a impactar en el detector mientras que el resto son desviados y filtrados.
- Las **Trampas Iónicas** funcionan bajo el mismo principio físico que los cuadropolos, pero la conformación en forma de cámara de las trampas iónicas permite confinar y acumular los iones que luego son liberados selectivamente.

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

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

Orbitrap (Figura 1.4 b) es un tipo de trampa iónica, relativamente reciente, desarrollado a finales de los años 90 del s.XX (Makarov, 2000) Consiste en un electrodo en un eje interno rodeado por un electrodo externo cilíndrico. Los iones

INTRODUCCIÓN

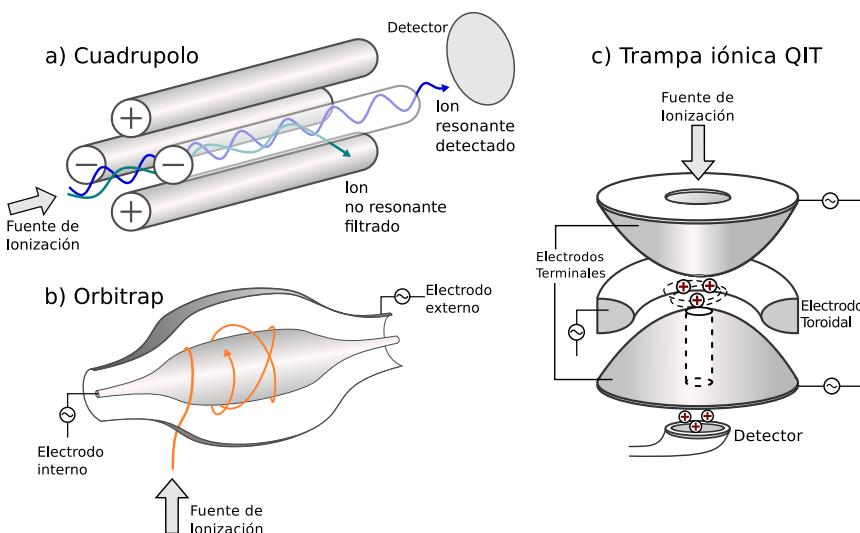


Figura 1.4: Algunos de los analizadores de masas más comunes son los de tipo Cuadrupolo (a), Orbitrap (b) y Trampa Iónica Tridimensional (c)

son introducidos tangencialmente desde la fuente de ionización y, al ajustar la diferencia de potencial, son atrapados en órbitas elípticas longitudinales, en las que la atracción hacia el eje interno es compensada por la fuerza centrífuga. La relación m/z se determina a partir de la frecuencia angular de la oscilación de los iones en torno al electrodo longitudinal interno.

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

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

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

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

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

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

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

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

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

INTRODUCCIÓN

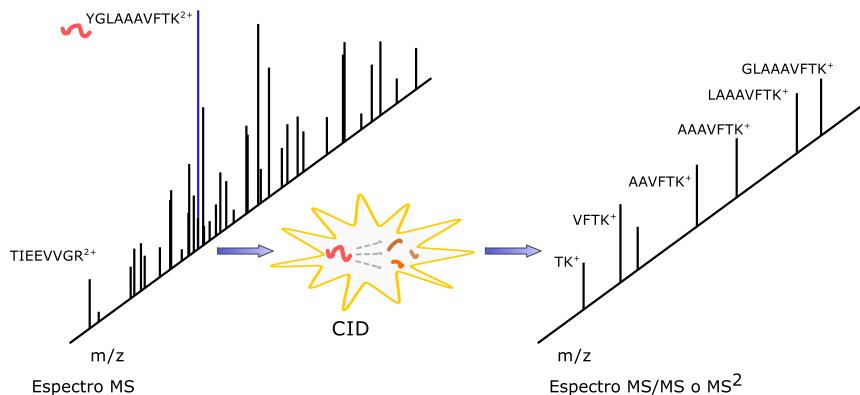


Figura 1.5: En la espectrometría de masas en Tándem se seleccionan péptidos precursores, generalmente los de mayor intensidad en el espectro MS¹, para ser fragmentados y generar el espectro MS/MS o MS²

1.2.3. Espectrometría de masas en Tándem. MS/MS

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

En la Espectrometría de Masas en Tándem, MS/MS (*Tandem Mass Spectrometry*) o MS², los péptidos, una vez ionizados y dentro del especlómetro, son sometidos a una fragmentación adicional. (Figura 1.5). Los péptidos se fragmentan, generando iones más pequeños lo que hace que el patrón de fragmentación sea más específico de la secuencia original. Esto aumenta el poder de resolución del análisis, ya que permite distinguir péptidos que, intactos,

tienen masas muy similares, pero cuyos patrones de fragmentación MS/MS son diferentes. Esto posibilita, además, partir de muestras con mezclas de proteínas más complejas incrementando así el rendimiento del experimento.

El proceso que conduce a la adquisición de espectros conlleva varias etapas. En primer lugar el instrumento escanea todos los péptidos ionizados introducidos en el espectrómetro y registra los llamados espectros MS¹, valores m/z y sus correspondientes intensidades para cada ion. A continuación, en función de la intensidad registrada en MS¹, se seleccionan y aislan algunos de estos iones -precursores- para ser fragmentados en péptidos más pequeños -fragmentos- en el interior del espectrómetro. El espectro MS² adquirido o espectro de fragmentación, registra los valores m/z e intensidades de los fragmentos de cada uno de los péptidos precursores aislados y fragmentados. El patrón de fragmentación codificado en los espectros MS² contiene la información necesaria para deducir la secuencia aminoacídica del péptido que lo origina.

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

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

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

Técnicas de fragmentación

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

INTRODUCCIÓN

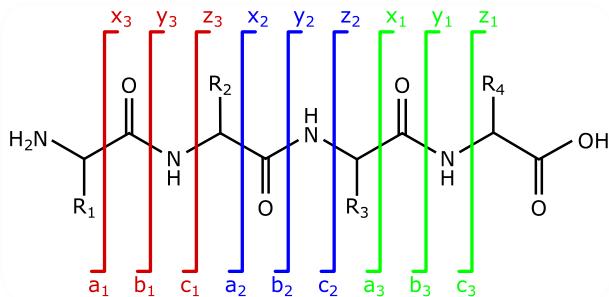


Figura 1.6: Nomenclatura de Roepstorff para los fragmentos en MS/MS

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

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

1.3. Digestión de proteínas en péptidos

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

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

La respuesta a estas preguntas, revisada en (Steen and Mann, 2004), tiene que ver, sobre

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

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

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

1.4. Proteómica en gel

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

En la Electroforesis en Gel de PoliAcrilamida Monodimensional, 1D-PAGE (*Monodimensional PolyAcrylamide Gel Electrophoresis*), las proteínas se separan en función de su peso molecular, las más pequeñas avanzan más en el gel. En la Electroforesis en Gel de PoliAcrilamida Bidimensional, 2D-PAGE (*Bidimensional PolyAcrylamide Gel Electrophoresis*), (Klose, 1975; O'Farrell, 1975), las proteínas se separan en una primera dimensión (sobre una tira con un gradiente de pH inmovilizado) en función de su punto isoelectrónico para posteriormente, aplicar la segunda dimensión, equivalente a 1DPAGE.

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

INTRODUCCIÓN

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 o muy hidrofóbicas no son fácilmente observables (Gygi et al., 2000), o una limitación en el número de proteínas identificables a partir de un gel que difícilmente puede superar el millar (Klose and Kobalz, 1995).

Sin embargo, este tipo de estudios sigue teniendo un nicho en la Proteómica actual (Rogowska-Wrzesinska et al., 2013). Notablemente, permite la visualización, identificación y cuantificación de proteínas intactas. La particular capacidad de la Proteómica en gel para separar proteínas con pequeños cambios en sus puntos isoeléctricos, *pI*, permite discernir entre isoformas de proteínas, o versiones de la misma proteína con PTM, lo que difícilmente se puede conseguir con otro tipo de aproximaciones.

1.4.1. Huella Peptídica

La Huella Peptídica de una proteína se refiere al hecho de que el patrón de fragmentación de una proteína en los péptidos que la constituyen utilizando una enzima proteolítica determinada, es muy específico de la proteína originaria (siempre y cuando se conozca el patrón de corte de la enzima, como es el caso de la tripsina) de forma que el espectro que generan puede ser utilizado para 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, para obtener una huella peptídica se requiere que la proteína se encuentre previamente aislada, generalmente a partir de una *mancha* o *spot* proteico de 2D-PAGE

La técnica de la Huella de Masas Peptídicas, PMF (*Peptide Mass Fingerprint*), desarrollado a principios de los años 90 por varios grupos independientemente (Pappin et al., 1993; Henzel et al., 1993; ?) se lleva a cabo generalmente por espectrometría de masas MALDI-TOF(ToF). Esto significa que, una vez obtenidos los péptidos correspondientes a la proteína del *spot*, éstos se sitúan en una matriz MALDI, donde son ionizados e introducidos en un analizador TOF. Una vez obtenido el espectro patrón de masas peptídicas, el proceso de análisis consiguiente es similar al que se hace en la Proteómica a gran escala o *shotgun*. Como se describe en las secciones siguientes, la identificación del péptido responsable del espectro se realiza utilizando un motor de búsqueda, que compara los valores de *m/z* del espectro obtenidos empíricamente con los valores de *m/z* calculados a partir de las secuencias de péptidos trípticos teóricos.

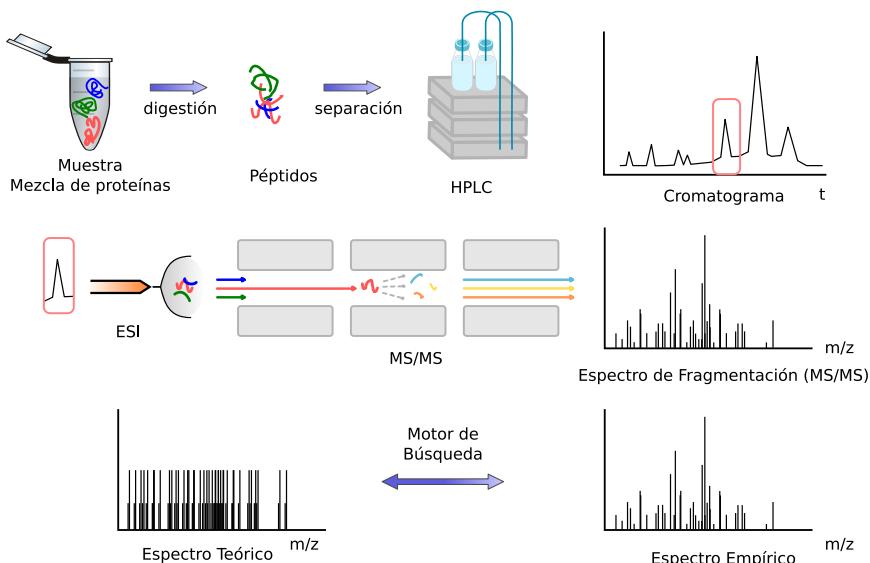


Figura 1.7: Etapas en un experimento de Proteómica *Shotgun*. En primer lugar la muestra de proteínas es digerida en péptidos. A continuación éstos son separados mediante cromatografía. A medida que eluyen, los péptidos separados, son introducidos en el espectrómetro de masas que genera un espectro de fragmentación MS/MS. Por último el motor de búsqueda compara los espectros empíricos con espectros teóricos asignando una puntuación.

1.5. Proteómica a gran escala o de alto rendimiento

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

1.5.1. Separación de péptidos y proteínas sin gel

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

INTRODUCCIÓN

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

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

Cromatografía Líquida de Alto Rendimiento. HPLC

Pero el fraccionamiento más importante se hace a nivel de péptido, trás la digestión de las proteínas, por HPLC. El funcionamiento básico general en HPLC 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 Cromatografía Líquida de Alto Rendimiento en Fase Reversa, RP-HPLC (*Reverse Phase High Performance Liquid Chromatography*). En ella los analitos de la muestra se separan en base a su carácter hidrofóbico. La fase estacionaria, apolar, está compuesta por unas micro-esferas de sílice cubiertas de cadenas alquilo con 18 átomos de C (C18). Un gradiente en que el solvente orgánico aumente gradualmente y en proporción inversa al componente acuoso, provoca que los analitos más polares eluyan primero integrados en la fracción acuosa cuando hay una mayor proporción de ésta, mientras que los más hidrofóbicos son retenidos durante más tiempo. Además, la cromatografía, usando una terminología similar a la usada para la separación de proteínas en gel, también puede ser multi-dimensional. La Tecnología Multidimensional de Identificación de Proteínas MudPIT (*Multidimensional Protein Identification Technology*) (Wolters et al., 2001) integra una primera dimensión de separa-

ción usando una columna de intercambio catiónico SCX (*Strong Cation Exchange*) y segunda dimensión consistente en una cromatografía en Fase Reversa.

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

Tras la adquisición experimental de los espectros, el paso siguiente en un experimento de Proteómica *shotgun* implica el análisis computacional de esos espectros cuyo objetivo final es la obtención de una lista de proteínas que presumiblemente se encuentran en la muestra. Este análisis computacional, a su vez, consta de varios procesos secuenciales, principalmente la asignación de secuencias peptídicas a cada espectro (sección 1.6), la inferencia de las proteínas a partir de esos péptidos (sección 1.8) y una evaluación estadística que aporta medidas de fiabilidad a la identificación (sección 1.7).

1.6. Asignación Péptido-Espectro

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

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

1.6.1. Búsqueda en bases de datos de secuencias

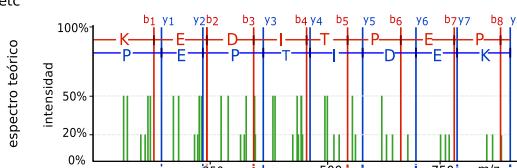
La búsqueda utilizando bases de datos de secuencias es el principal y más extendido método de asignación de una secuencia peptídica a un espectro MS/MS (figura 1.8). Existen una gran variedad de herramientas computacionales llamadas motores de búsqueda diseñadas

INTRODUCCIÓN

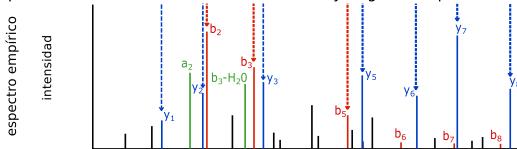
Interpretación automática usando espectros teóricos. Principios básicos

1 - Dado el conocimiento previo sobre:
·Secuencias en BBDD de proteínas
·Probabilidad de que cada tipo de íon se encuentre en un espectro MS/MS:

2 - Se puede crear un espectro teórico para una lista de péptidos candidatos elaborada conociendo:
·Enzima proteolítica utilizada
·Rango de tolerancia de masas
·etc



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



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

Figura 1.8: Interpretación automática de espectros MS/MS. Gracias a la información disponible sobre la presencia de los diferentes tipos de iones se puede elaborar una lista de péptidos candidatos que han generado el espectro empírico y para cada uno de ellos comparar y evaluar su similitud.

para realizar esta tarea.

Los motores de búsqueda 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 es posible obtener gracias a diversas fuentes de información disponible como el patrón de corte de la enzima proteolítica utilizada, los valores m/z de los fragmentos que se producirían a partir de los péptidos, la frecuencia estimada de cada tipo de fragmento y las secuencias de las proteínas en una base de datos de referencia. En el proceso, el espacio de búsqueda se acota mediante la selección de una lista de péptidos posibles, (candidatos que cumplen unos criterios determinados), que posiblemente han generado el espectro MS/MS y a continuación se ordenan utilizando una puntuación función del grado de similitud entre el espectro empírico y el teórico.

Elaboración de una lista de péptidos candidatos

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

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

Otros parámetros que se proporcionan al *software* y que afectan notablemente a la creación de la lista de candidatos y por tanto también al coste computacional son, la selección de masa mono-isotópica o masa promedio (es decir, considerar que todos los átomos de C se encuentran en su forma ^{12}C o bien que exista una proporción variable de isótopos ^{13}C); el número de puntos de corte no efectuados permitidos dentro de la secuencia del precursor; la existencia de modificaciones post-traducionales y otras modificaciones permitidas (variables o fijas) que ocurren en el proceso experimental; y la selección del tipo de iones fragmento a buscar.

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

INTRODUCCIÓN



Figura 1.9: Estrategia básica de identificación. Para el conjunto seleccionado de péptidos candidatos se establece una correlación entre el espectro MS/MS empírico y el espectro generado teóricamente por cada uno de los candidatos.

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

Motores de búsqueda. Funciones de puntuación

Los motores de búsqueda se encargan de asignar a cada espectro empírico obtenido un péptido, el mejor candidato de una lista de los posibles péptidos que han generado ese espectro, con una cierta medida de puntuación función del grado de similitud entre espectro empírico y teórico. La estrategia general consiste en realizar una digestión teórica de las secuencias de proteínas de referencia, teniendo en cuenta los parámetros especificados. Así, para cada espectro observado, el motor de búsqueda recorre las secuencias en una base de datos (un archivo FASTA) seleccionando aquellos péptidos con valores *m/z* similares al del ion precursor en el espectro empírico y que se encuentran dentro del rango de tolerancia permitido obteniendo un espectro teórico para cada uno. A continuación se establece el grado de similitud de cada espectro adquirido con los espectros teóricos de cada uno de 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 (Eng et al., 1994) fue la primera herramienta descrita para correlacionar espectros MS/MS con secuencias de aminoácidos y actualmente sigue siendo uno de los programas más utilizados. Para cada espectro adquirido, SEQUEST calcula de manera independiente la puntuación de correlación (*cross-correlation Score*, $Xcorr$) para todos los candidatos con los que es comparado. En primer lugar se crea un espectro empírico procesado (espectro X) en el que los picos de baja intensidad son eliminados y el resto de valores m/z son redondeados al valor entero más próximo. Para cada candidato se crea un espectro teórico (espectro Y) usando unas reglas de fragmentación simplificadas. Entonces el valor $Xcorr$ es calculado usando la función de correlación $Corr(t)$ (el producto entre los vectores X e Y, con Y desplazado t unidades de masa respecto a X a lo largo del eje m/z) (a)

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

donde x_i e y_i representan los picos del espectro procesado y el espectro teórico respectivamente. 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 sustrayendo el valor medio de $Corr(t)$ en torno a $t = 0$, que representa una estimación de las coincidencias aleatorias entre picos (b). SEQUEST también proporciona un valor de puntuación adicional, ΔCn , que indica la diferencia entre el valor $Xcorr$ del mejor candidato y el del segundo mejor candidato. Ambos valores son por tanto indicativos de la calidad de cada PSM que será mejor cuanto más altas sean ambas puntuaciones.

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

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

INTRODUCCIÓN

b- e y-.

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

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

El motor de búsqueda OMSSA (*Open Mass Spectrometry Search Algorithm*) (Geer et al., 2004), al igual que X!Tandem, es de código abierto. En OMSSA, la puntuación es un reflejo del número de coincidencias entre iones del espectro experimental y el teórico sin tener en cuenta si los iones son de tipo *b*- o *y*-.

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

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

1.6.2. Otras estrategias de asignación péptido-espectro

Búsqueda basada en bibliotecas de espectros

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

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

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 (Lam et al., 2007) Además,

1.7. Evaluación estadística de las identificaciones de péptidos y proteínas

a los resultados obtenidos también se les puede aplicar los modelos de validación estadística desarrollados para las búsquedas basadas en secuencia.

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

Identificación por secuenciación *de novo*

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

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

1.7. Evaluación estadística de las identificaciones de péptidos y proteínas

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

1.7.1. Conceptos estadísticos básicos

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

INTRODUCCIÓN

términos globales, para todo un experimento, las puntuaciones de todos los PSM.

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

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

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

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

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

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

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

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

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

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

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

1.7.2. Puntuaciones basadas en distribuciones de espectro individual y promedio

Distribución de espectro individual

Generalmente los motores de búsqueda aportan varios tipos de puntuación para cada PSM. Un tipo de puntuaciones se refiere a la calidad de cada asignación en particular, evalúa el grado de similitud entre el espectro empírico y el péptido asignado, el mejor de la lista de candidatos. (*Xcorr* en SEQUEST, *hyperscore* en X!Tandem o *ionScore* en Mascot). Pero

1.7. Evaluación estadística de las identificaciones de péptidos y proteínas

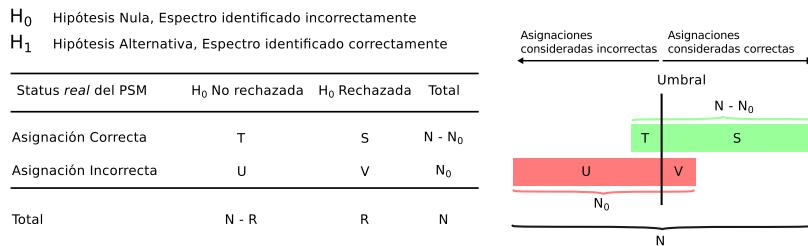


Figura 1.10: Tabla de contingencia. Contraste de hipótesis. La estrategia general consiste en realizar tests que permitan rechazar la hipótesis nula y por tanto tener un criterio estadístico para aceptar la hipótesis alternativa, es decir, afirmar que la asignación es correcta.

además, a veces, aportan una puntuación indicativa de la calidad del PSM en relación a otros, al segundo mejor candidato (ΔCn de SEQUEST); o con respecto a una población del resto de candidatos que obtuvieron puntuaciones inferiores utilizando los parámetros estadísticos *e*-valor y *p*-valor.

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

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

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

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

INTRODUCCIÓN

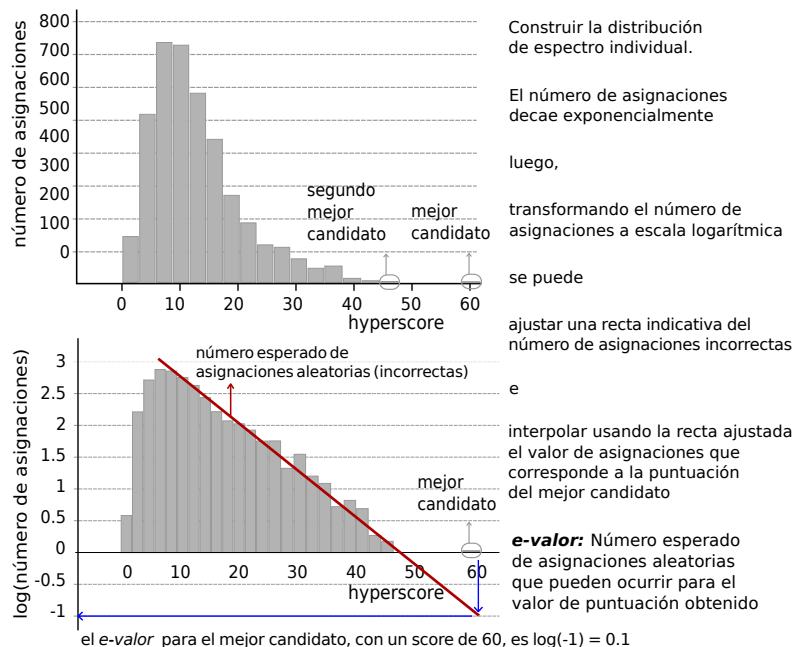


Figura 1.11: Para estimar el e-valor, X!Tandem convierte a escala logarítmica la distribución de espectro individual utilizando su puntuación (*hyperscore*) y a continuación, interpola, en la recta ajustada, (que representa el número de asignaciones aleatorias esperadas) el valor que correspondería a la mejor puntuación obtenida.

parámetros (Nesvizhskii, 2010).

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

Distribución promedio

En los experimentos *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

1.7. Evaluación estadística de las identificaciones de péptidos y proteínas

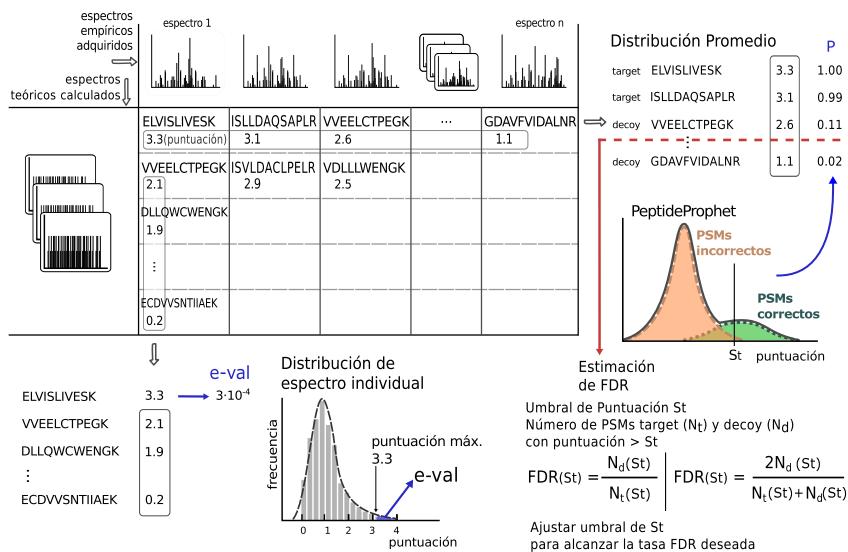


Figura 1.12: La distribución de espectro individual es la distribución de las puntuaciones de todos los péptidos candidatos comparados con el espectro empírico. La distribución promedio, es la distribución de las puntuaciones de los mejores candidatos para el total de todos los espectros empíricos. Ésta se puede considerar una población mixta compuesta por una mezcla de subpoblación de asignaciones incorrectas y otra subpoblación de asignaciones correctas.

podrían ocurrir PSM con *p*-valores igualmente bajos sólo 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 (Abdi, 2007), que simplemente divide el *p*-valor por el número de veces que se repite el test. Así para un PSM con *p*-valor = 0,05 en un experimento en el que hay otros 10.000 PSM, el *p*-valor original habría de readjustarse a $0,05/10.000 = 5 \cdot 10^{-6}$.

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

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

INTRODUCCIÓN

e-valores obtenidos individualmente.

1.7.3. 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)* (Benjamini and Hochberg, 1995). Básicamente, el concepto de tasa FDR se refiere a la proporción de PSM incorrectos que se aceptan en todo el conjunto de PSM de un experimento para un umbral de puntuación (o de parámetro estadístico como el *p*-valor) fijado.

Para la estimación de la tasa FDR (Figura 1.12), la estrategia utilizada consiste esencialmente en utilizar una base de datos llamada *señuelo* o *decoy* (Figura 1.13) (Elias and Gygi, 2007). Es una aproximación sencilla pero efectiva que requiere que los espectros MS/MS sean comparados con espectros teóricos derivados de secuencias *señuelo*, que pueden ser generadas de varias formas pero que, en cualquier caso, son secuencias que no existen, no corresponden a ninguna proteína. 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 (*secuencias reales - secuencias señuelo*) del doble de tamaño que la original. O bien se pueden realizar dos búsquedas consecutivas, una utilizando la base de datos de secuencias *reales* y otra a continuación utilizando la de secuencias *señuelo*.

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

Una vez establecida esta hipótesis nula, la estrategia asume una idea básica central: la

1.7. Evaluación estadística de las identificaciones de péptidos y proteínas

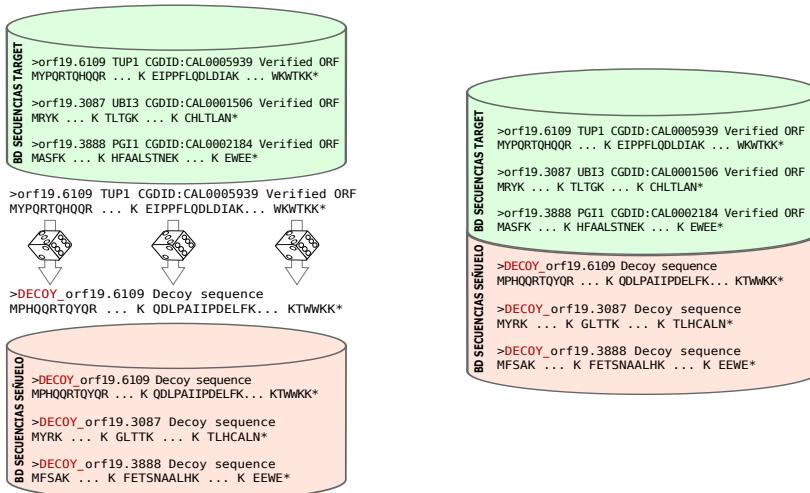


Figura 1.13: Una estrategia comúnmente empleada en la construcción de bases de datos señouelo consiste en barajar la secuencia aminoacídica, péptido a péptido (para así conservar el tamaño medio de los péptidos), de cada una de las proteínas en la base de datos original, para obtener así una base de datos señouelo, y a continuación concatenarlas.

frecuencia con que los espectros MS/MS son asignados a secuencias *señouelo* sigue la misma distribución que la frecuencia con que los espectros son asignados incorrectamente a secuencias *reales*.

Así, de forma general y dado que las bases de datos de secuencias *reales* y *señouelo* tienen el mismo tamaño, el número de PSM incorrectos o Falsos Positivos (N_{inc} , aquellos espectros a los que se ha asignado incorrectamente una secuencia *real*) puede ser considerado equivalente al número de PSM *señouelo* (N_d , espectros a los que se ha asignado una secuencia *señouelo*). Con esto se puede estimar la tasa FDR como N_d/N_t , esto es, la proporción de PSM *señouelo*, N_d como sustituto conocido de N_{inc} , entre el total de secuencias *reales* 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 *señouelo*, 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 com-

INTRODUCCIÓN

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

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

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

1.7.4. Modelos mixtos de probabilidad. Probabilidad Posterior

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

PeptideProphet (Keller et al., 2002) es un algoritmo de post-procesamiento (empleado después de que el motor de búsqueda haya establecido una lista de PSM), el primero en implementar este tipo de análisis, que permite estimar la confianza en los péptidos identificados aportando un valor de probabilidad posterior.

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

Primero, PeptideProphet recalcula las puntuaciones. Mediante análisis discriminante, las

1.7. Evaluación estadística de las identificaciones de péptidos y proteínas

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

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

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

Y como ejemplo específico para el caso de SEQUEST:

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

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

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

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

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

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

INTRODUCCIÓN

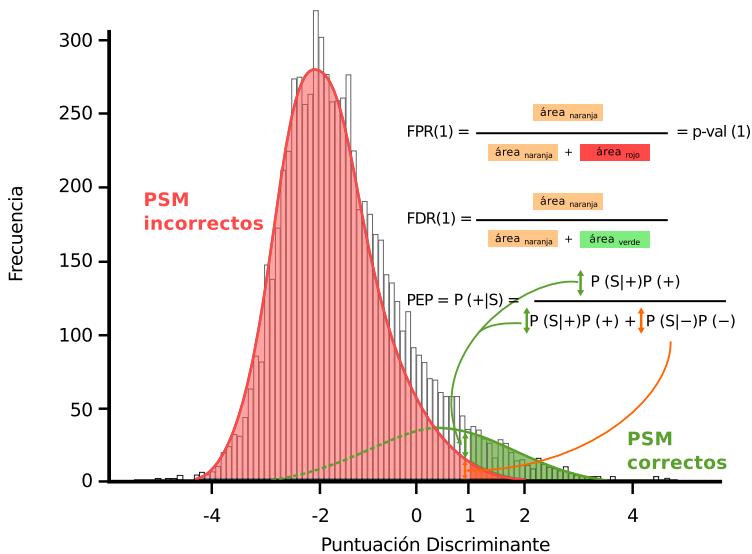


Figura 1.14: PeptideProphet considera la distribución promedio de PSMs como una mezcla de una subpoblación de PSM correctos (distribución Normal, en verde) y otra de PSM incorrectos (distribución Gamma, en rojo). Una vez inferidas estas distribuciones mediante el algoritmo *EM*, la probabilidad de un PSM en particular puede calcularse mediante el teorema de Bayes como la probabilidad de tener una puntuación siendo correcto con respecto a la probabilidad de tener esa puntuación.

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

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

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

coincide con la FDR local.

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

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

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

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

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

En un principio, cuando fue implementado (Keller et al., 2002), este modelo mixto para el cálculo de probabilidades no hacía uso de búsquedas realizadas usando la estrategia *target-señuelo* pues no se había generalizado aún. Cuando comenzaron a emplearse este tipo de búsquedas PeptideProphet incorporó esta información haciendo que las puntuaciones correspondientes a péptidos *señuelo* sólamente puedan contribuir a la estimación de los parámetros que describen la distribución *Gamma* de PSM incorrectos. Esto permitió redefinir una versión semisupervisada del algoritmo (Choi and Nesvizhskii, 2008) en el sentido de que la clase (PSM correctos e incorrectos) pasa a ser conocida para algunas pero no todas las asignaciones.

1.8. Inferencia de proteínas a partir de péptidos

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

Uno de los principales motivos que complican la inferencia de las proteínas se refiere a la pérdida de la correspondencia péptido-proteína como consecuencia fundamentalmente de la detección de péptidos compartidos o *degenerados* cuyas secuencias están en diferentes proteínas. Esta dificultad, descrita como *el problema de la inferencia de proteínas* en (Nesvizhskii and Aebersold, 2005), a su vez se deriva de varios posibles escenarios. En algunos casos el procesamiento alternativo de intrones provoca la existencia de isoformas de una proteína en la muestra. De éstas, sólo algunas tienen en su secuencia péptidos trípticos exclusivos que permitan, en caso de ser detectados, concluir la presencia de la proteína fehacientemente. En otros casos, proteínas diferentes procedentes de una familia de genes (parálogos) poseen una alta homología de secuencia. Frecuentemente, a pesar de detectar un cierto número de

INTRODUCCIÓN

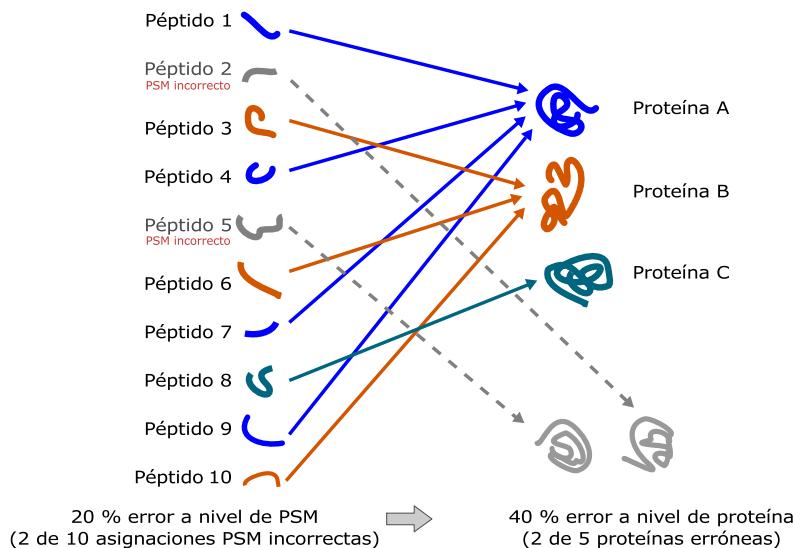


Figura 1.15: Agrupamiento no aleatorio de péptidos en proteínas. Los péptidos correctamente asignados generalmente se agrupan formando parte de una proteína, sin embargo esto no ocurre para las asignaciones incorrectas, cada PSM incorrecto corresponde a una proteína diferente y se traduce por tanto en una identificación de proteína errónea. (1 PSM incorrecto, 1 proteína incorrecta).

péptidos, no se puede indicar la presencia de ninguna de las proteínas de la familia en particular.

En función de la presencia de estos péptidos compartidos y de péptidos exclusivos de cada proteína se han descrito métodos y nomenclaturas adecuadas para definir como *identificada* una proteína o lista de proteínas (Nesvizhskii and Aebersold, 2005; Prieto et al., 2012)

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

Otra de las dificultades en la inferencia de proteínas es el fenómeno consistente en el agrupamiento dirigido de péptidos en sus correspondientes proteínas, lo que provoca una amplificación de las tasas de error. Como se observa en la figura 1.15, mientras que las identificaciones de péptidos correctas tienden a agruparse en un número pequeño de proteínas, los

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

INTRODUCCIÓN

péptidos incorrectos, puesto que proceden de asignaciones aleatorias a entradas en la base de datos, suponen una proteína errónea por cada péptido.

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

$$NSP_i = \sum_{j \neq i} p_j$$
$$p'_i = \frac{p_i f_1(NSP_i)}{p_i f_1(NSP_i) + (1 - p_i) f_0(NSP_i)}$$
 (1.14)

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

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

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

Algunas herramientas bioinformáticas de post-procesamiento de resultados de identificaciones como PeptideProphet y ProteinProphet han demostrado ser de gran utilidad proporcionando un medio de calcular probabilidades y tasas de error de forma muy precisa. Sin

INTRODUCCIÓN

embargo, la modelización que emplean a veces no resulta muy eficiente, especialmente en casos de listas de espectros e identificaciones muy grandes (Reiter et al., 2009).

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

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

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

1.10. Proteómica dirigida. SRM/MRM

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

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

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

Básicamente, *SRM* proporciona unas propiedades muy interesantes en 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 (Gerber et al., 2003), 2004 (Kuhn et al., 2004).

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

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

1.11. Repositorios públicos de Proteómica online

Los resultados generados en experimentos de Proteómica han de ser convenientemente mostrados y compartidos con la comunidad científica. Algunos de los repositorios más populares que almacenan y permiten visualizar resultados de experimentos de Proteómica son PRIDE y PeptideAtlas.

INTRODUCCIÓN

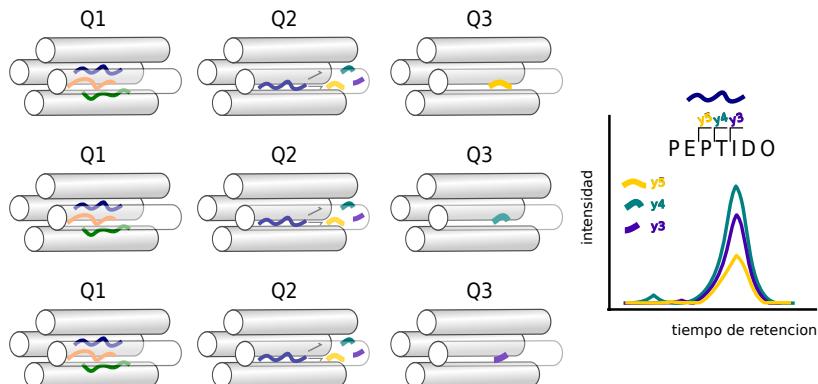


Figura 1.16: Adquisición y reconstrucción de la señal en un experimento SRM

1.11.1. PRIDE

La base de datos PRIDE (*Protein Identifications Database*), creada en el Instituto Bioinformático Europeo en Inglaterra (Martens et al., 2005) es el repositorio más extenso de datos de espectrometría de masas. Almacena los resultados originales enviados por los investigadores usando para ello su propio formato PRIDE XML. Además se han desarrollado herramientas que facilitan la creación y el envío de los resultados en este formato.

1.11.2. PeptideAtlas

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

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

En el proceso de análisis de datos que sigue a la adquisición experimental de espectros se requiere un uso intensivo de *software*, desde la asignación de secuencias peptídicas a los espectros hasta la elaboración de listas de proteínas 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. En este sentido la iniciativa HUPO-PSI ha adquirido un papel muy importante en la elaboración y difusión de formatos abiertos que puedan servir de estándar para toda la comunidad

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

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

Este es un tipo de formatos muy diverso. Depende básicamente, de la forma en que se detectan y recogen 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 instrumentos registran los espectros en modo continuo de forma predefinida, pero frecuentemente son sometidos a un procesamiento por un algoritmo que extrae los picos detectados como parejas de valores *m/z* e intensidad. Esto se denomina adquisición de *datos centroide*

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

Sin embargo, el hecho de que estos formatos sean codificados en binario junto con la disponibilidad de las librerías de lectura proporcionadas por los fabricantes restringida

INTRODUCCIÓN

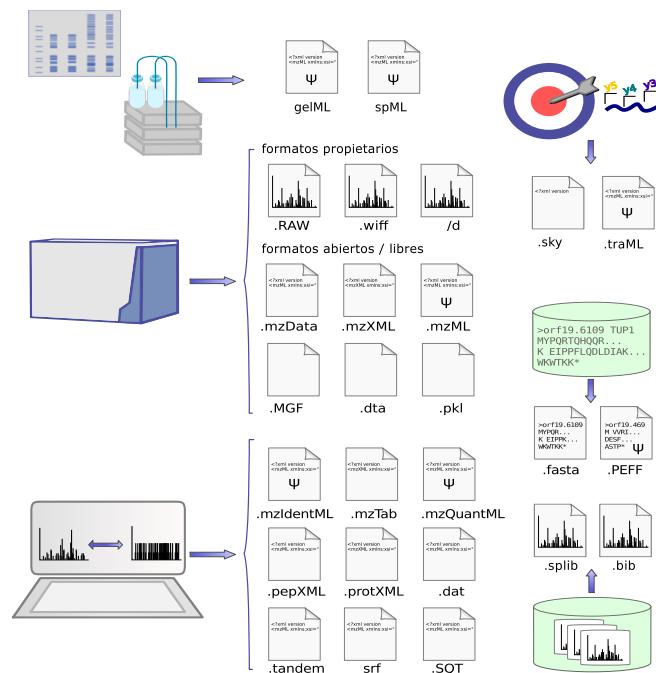


Figura 1.17: Visión general de formatos comúnmente usados en cada etapa de un experimento de Proteómica. Los formatos adoptados como estándar por HUPO-PSI contienen el símbolo Ψ

únicamente al sistema operativo MS Windows frenó el desarrollo de nuevas herramientas de lectura y manipulación de este tipo de archivos. En ese contexto, aparecieron los primeros formatos de texto (XML) para codificar toda la información de salida de MS, mzXML (Pedrioli et al., 2004) y mzData que posteriormente fueron unificados en el estándar HUPO-PSI mzML (Deutsch, 2010).

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

- Formatos que recogen el resultado de las búsquedas

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

INTRODUCCIÓN

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

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

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

■ *Formatos que almacenan bibliotecas de espectros*

Los motores de búsqueda que usan bibliotecas de espectros, como SpectraST (parte de TPP) (Lam et al., 2007) requieren generalmente un formato que contenga espectros consenso, una combinación de los espectros y el péptido que se les ha asignado, así como otras anotaciones y metadatos. El Instituto Nacional para Estándares y Tecnología americano, NIST (*National Institute for Standards and Technology*) distribuye bibliotecas de espectros en formato *.msp*. SpectraST produce el formato *.splib*; y X!Hunter y BiblioSpec *ASL* y *.blib* respectivamente.

■ *Formatos que almacenan secuencias*

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

■ *Formatos específicos para Proteómica dirigida*

En Proteómica dirigida, podemos encontrar dos tipos de formatos. Entre los que se usan como fuente de entrada de información, para indicar al espectrómetro las listas

INTRODUCCIÓN

de transiciones, es decir, que precursores y fragmentos ha de monitorizar, los más empleados son *.sky*, empleado por el programa Skyline; y el estándar HUPO-PSI *TraML* (Deutsch et al., 2012). Para los resultados de análisis por SRM, el programa Skyline usa su propio formato, *.skyd*, basado en XML y otros programas como mProphet (Reiter et al., 2011) usan sus propios formatos de texto separado por tabulador.

1.13. *Candida albicans* como organismo modelo

Candida albicans es un hongo patógeno oportunista que se encuentra comúnmente como residente comensal, inocuo, en las mucosas gastrointestinal y urogenital en un alto porcentaje de la población (Calderone, 2012). Sin embargo, su cualidad de patógeno oportunista implica que, en ocasiones, propiciadas generalmente por un sistema inmune debilitado en el hospedador, puede proliferar y diseminarse provocando infecciones, candidiasis, de gravedad variable, desde afecções mucocutáneas leves hasta infecciones sistémicas severas que pueden incluso llegar a ser letales. En Estados Unidos especies del género *Candida* suponen la cuarta causa más común de infecciones nosocomiales con tasas de mortalidad de hasta el 50 % en el caso de infecciones sistémicas (Pfaller and Diekema, 2010). Los principales factores de virulencia con los que cuenta *Candida albicans* para proliferar y diseminarse causando infecciones son su polimorfismo, su capacidad de producir adhesinas e invasinas, y la formación de biopelículas (Mayer et al., 2013).

El polimorfismo consiste en la capacidad de transformar su morfología. Las morfologías más comunes (además de otras como las pseudohifas y las clamidosporas) son la clásica forma ovalada levaduriforme, adecuada para la diseminación a través de los vasos sanguíneos, y los filamentos o hifas que permiten penetrar e invadir tejidos (Berman and Sudbery, 2002; Jacobsen et al., 2012).

La producción de adhesinas, proteínas especializadas en la adhesión a las superficies abióticas o de otras células es otro importante factor de virulencia. Las aglutininas de la familia ALS que incluyen proteínas con anclaje GPI como Als3 (Phan et al., 2007; de Groot et al., 2013) y otras como la también anclada mediante GPI y asociada a hifas Hwp1 (Sundstrom et al., 2002) son algunas de las adhesinas más estudiadas.

La formación de biopelículas sobre sustratos abióticos como catéteres o prótesis dentales o sobre la superficie celular en mucosas también supone un factor de virulencia crítico. Se ha descrito que biopelículas maduras adquieren una mayor resistencia a antifúngicos y a la acción del sistema inmune del hospedador en comparación con células planctónicas (Fanning and Mitchell, 2012).

Desde el punto de vista de la Proteómica, se han realizado muy diversos estudios usando el modelo de *Candida albicans*. En este sentido los experimentos proteómicos se han enfocado principalmente en el estudio de los rasgos patógenos. Existen estudios enfocados al

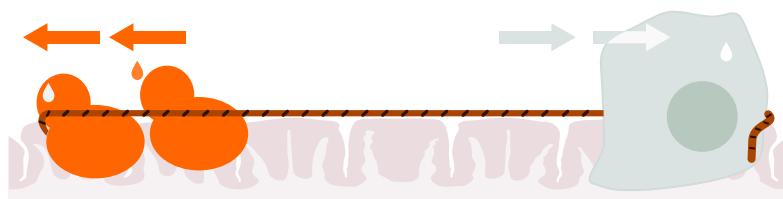


Figura 1.18: La población comensal de células de *Candida albicans* y células del sistema inmune como macrófagos, presentes en las mucosas de los tractos gastrointestinal y genitourinario se encuentran en un *status* de equilibrio en condiciones normales.

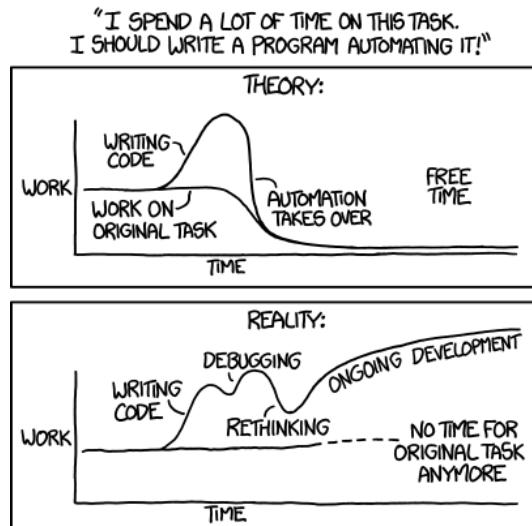
estudio del proteoma de la pared celular (Castillo et al., 2008), estudios de proteómica de los factores de virulencia (Pitarch et al., 2006a), de la respuesta serológica (Pitarch et al., 2009) y también estudios sobre la resistencia a antifúngicos (Hoehamer et al., 2010).

OBJETIVOS

Objetivos

- Desarrollo de una aplicación web, denominada Proteopathogen (v 1.0), respaldada por una base de datos para recoger, almacenar y analizar resultados de identificación de péptidos y proteínas procedentes de experimentos de proteómica relacionados con el estudio de la interacción hospedador-patógeno usando el modelo de *Candida albicans*
- Adopción del formato estándar de identificaciones de péptidos y proteínas mzIdentML como fuente de datos para la base de datos y aplicación web Proteopathogen (v 2.0)
- Recopilación de resultados de espectrometría de masas y creación de un PeptideAtlas, o Atlas Peptídico, de *Candida albicans*
- Reanálisis de los datos existentes en el PeptideAtlas original usando una nueva base de datos con secuencias específicas de alelo, un procesamiento multi-algoritmo y resultados de nuevos experimentos diseñados *ad hoc* para incrementar la cobertura del proteoma.

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



xkcd, Randall Munroe

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

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

Proteomics 2009, 9, 4664-4668

Abstract

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>

PROTEOPATHOGEN DATABASE TO STUDY *C. albicans* - HOST INTERACTION

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 (Calderone, 2012).

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 (Fernández-Arenas et al., 2007; Martínez-Solano et al., 2006; Pitarch et al., 2006a,b).

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 (Monteoliva and Albar, 2004). Existing public repositories for management and storage of proteomic data such as World 2-D PAGE (Hoogland et al., 2008), the Proteome Database System for Microbial Research 2-D PAGE (Pleissner et al., 2004), or PRIDE (Martens et al., 2005); and fungi-specific databases such as BioBase MycoPathPD (Csank et al., 2002), Candida Genome Database (CGD) (Arnaud et al., 2005) or Candida DB (Rossignol et al., 2008) are very popular and useful tools. However, none of them deals with proteomic experimental 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 (Fernández-Arenas et al., 2007; Martínez-Solano et al., 2006), 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 (Cabezón et al., 2009). It compiles a set of experiments aimed at extraction and identification of membrane pro-

PROTEOPATHOGEN DATABASE TO STUDY *C. albicans* - HOST INTERACTION

| References | Description of experimental approach | Species | #Protein identifications |
|---------------------------------|---|---------------------|--------------------------|
| (Fernández-Arenas et al., 2007) | <i>C. albicans</i> differentially expressed proteins after 3 h interaction with RAW 264.7 murine macrophages. 2-D silver-stained gel. MS/MS (MALDI/TOF-TOF) | <i>C. albicans</i> | 66 |
| (Martínez-Solano et al., 2006) | Proteins identified from cytoplasmic extracts of RAW 264.7 cells after 45 min interaction with <i>C. albicans</i> | <i>Mus musculus</i> | 38 |
| (Cabezón et al., 2009) | Identification of Glycosyl phosphatidyl inositol (GPI)-anchored membrane proteins Identification of membrane proteins | <i>C. albicans</i> | 292 |
| | | | 1273 |

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

teins 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 (Plaine et al., 2008).

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 (Arnaud et al., 2005), whereas in the case of murine macrophage proteins, the equivalent information was obtained from UniProt KnowledgeBase (The Uniprot Consortium, 2008) and the Mouse Genome Database (Bult et al., 2008). Additionally, pathways annotations were retrieved from Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database (Kanehisa et al., 2007) and structure information from the Protein Data Bank (PDB) (Berman et al., 2000).

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.

PROTEOPATHOGEN DATABASE TO STUDY *C. albicans* - HOST INTERACTION

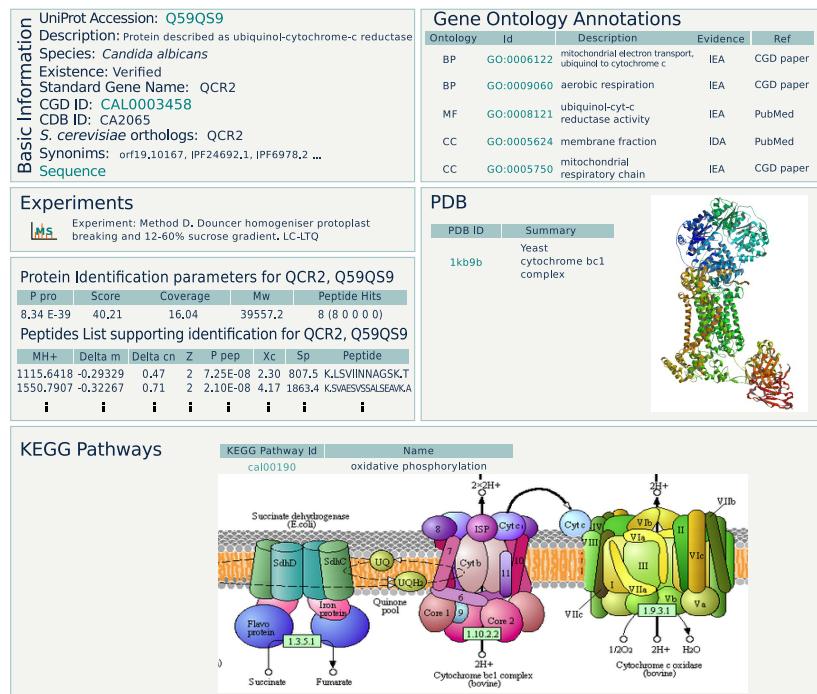


Figure 1. Use case: Search for *C. albicans* ubiquinol-cytochrome-c reductase QCR2. The different sections in the result comprise information on protein description and identifiers, experiments in which it has been identified, GO annotation, KEGG and CGD pathway annotation and structural information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

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.

PROTEOPATHOGEN DATABASE TO STUDY *C. albicans* - HOST INTERACTION

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, 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 and identification-supporting peptides lists. These data are subject to revision prior to eventual insertion into Proteo-pathogen 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 Proteo-pathogen 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 (Pitarch et al., 2004).

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.

PROTEOPATHOGEN DATABASE TO STUDY *C. albicans* - HOST INTERACTION

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 12-60sucrose 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 (Insenser et al., 2006), 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.

References

- Arnaud, M. B., Costanzo, M. C., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., and Sherlock, G. (2005), The Candida Genome Database (CGD), a community resource for *Candida albicans* gene and protein information., Nucleic acids research, 33(Database issue), D358-63.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000), The Protein Data Bank., Nucleic acids research,

PROTEOPATHOGEN DATABASE TO STUDY *C. albicans* - HOST INTERACTION

28(1), 235-42.

Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., and Blake, J. A. (2008), The Mouse Genome Database (MGD): mouse biology and model systems., Nucleic acids research, 36(Database issue), D724-8.

Cabezón, V., Llama-Palacios, A., Nombela, C., Monteoliva, L., and Gil, C. (2009), Analysis of *Candida albicans* plasma membrane proteome., Proteomics, 9(20), 4770-86.

Calderone, R. (2012), Candida and candidiasis. ASM Press, Washington DC.

Csank, C., Costanzo, M. C., Hirschman, J., Hodges, P., Kranz, J. E., Mangan, M., O'Neill, K., Robertson, L. S., Skrzypek, M. S., Brooks, J., and Garrels, J. I. (2002), Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD)., Methods in enzymology, 350, 347-73.

Fernández-Arenas, E., Cabezón, V., Bermejo, C., Arroyo, J., Nombela, C., Diez-Orejas, R., and Gil, C. (2007), Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction, Molecular & cellular proteomics : MCP, 6(3), 460-478.

Hoogland, C., Mostaguir, K., Appel, R. D., and Lisacek, F. (2008), The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server., Journal of proteomics, 71(2), 245-8.

Insenser, M., Nombela, C., Molero, G., and Gil, C. (2006), Proteomic analysis of detergent - resistant membranes from *Candida albicans*., Proteomics, 6 Suppl 1, S74-81.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2007), KEGG for linking genomes to life and the environment, Nucleic Acids Research, 36(Database), D480-D484.

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Van-dekerckhove, J., and Apweiler, R. (2005), PRIDE: the proteomics identifications database., Proteomics, 5(13), 3537-45.

Martínez-Solano, L., Nombela, C., Molero, G., and Gil, C. (2006), Differential protein expression of murine macrophages upon interaction with *Candida albicans*, Proteomics, 6 Suppl 1, S133-S144.

PROTEOPATHOGEN DATABASE TO STUDY *C. albicans* - HOST INTERACTION

- Monteoliva, L. and Albar, J. P. (2004), Differential proteomics: an overview of gel and non-gel based approaches., *Briefings in functional genomics & proteomics*, 3(3), 220-39.
- Pitarch, A., Abian, J., Carrascal, M., Sánchez, M., Nombela, C., and Gil, C. (2004), Proteomics-based identification of novel *Candida albicans* antigens for diagnosis of systemic candidiasis in patients with underlying hematological malignancies., *Proteomics*, 4(10), 3084-106.
- Pitarch, A., Nombela, C., and Gil, C. (2006a), *Candida albicans* biology and pathogenicity: insights from proteomics., *Methods of biochemical analysis*, 49, 285-330.
- Pitarch, A., Nombela, C., and Gil, C. (2006b), Contributions of proteomics to diagnosis, treatment, and prevention of candidiasis., *Methods of biochemical analysis*, 49, 331-61.
- Plaine, A., Walker, L., Da Costa, G., Mora-Montes, H. M., McKinnon, A., Gow, N. A. R., Gaillardin, C., Munro, C. A., and Richard, M. L. (2008), Functional analysis of *Candida albicans* GPI-anchored proteins: roles in cell wall integrity and caspofungin sensitivity., *Fungal genetics and biology : FG & B*, 45(10), 1404-14.
- Pleissner, K.-P., Eifert, T., Buettner, S., Schmidt, F., Boehme, M., Meyer, T. F., Kauffmann, S. H. E., and Jungblut, P. R. (2004), Web-accessible proteome databases for microbial research., *Proteomics*, 4(5), 1305-13.
- Rossignol, T., Lechat, P., Cuomo, C., Zeng, Q., Moszer, I., and DÉnfert, C. (2008), *CandidaDB*: a multi-genome database for *Candida* species and related *Saccharomycotina*., *Nucleic acids research*, 36(Database issue), D557-61.
- The Uniprot Consortium (2008), The universal protein resource (UniProt)., *Nucleic acids research*, 36(Database issue), D190-5.

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

Vital Vialas, Concha Gil

EuPA Open Proteomics 2015, 8, 22-27

Abstract

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

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

Introduction

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

Proteomics studies have been addressed to study this commensal to pathogenic transition by approaching the dimorphic, yeast form to hyphal form switch (Monteoliva et al., 2011; Gow and van de Veerdonk, 2011), by specifically aiming at the study of some other clinically relevant biological processes such as apoptosis (Madeo et al., 2004; Fernández-Arenas et al., 2007; Ramsdale, 2008) or biofilm formation (Vialás et al., 2012); or targeting sets of proteins that interact first with the host like surface exposed and secreted proteins (Vialás et al., 2012; Gil-Bona et al., 2014).

However, until recently, the resulting proteomics identification datasets were sparse and disseminated. The Proteopathogen database (Vialás et al., 2009) was the first public online proteomics data repository specifically focused on experiments aimed at the study of *C. albicans* and other fungal species pathogenic traits. Since no standard format for peptide and protein identification results was available, Proteopathogen was developed to compile and display identification lists in different tabulated text formats depending on the software used to generate and process the results.

At that time, the HUPO - Proteomics Standards Initiative (PSI) already had a trajectory striving to highlight the importance of standardization and providing formats that would comply with MIAPE (Minimum Information About a Proteomics Experiment) guidelines as reviewed in ref (Martínez-Bartolomé et al., 2014). Some *de facto* standard formats existed like mzXML and pepXML (Deutsch, 2012), but the advent, years later, of the HUPO-PSI approved formats for mass spectrometry output data (Martens et al., 2011) and for identification results (Jones et al., 2012) among others, surfaced the efforts and claims by the community to finally adopt formats to facilitate data comparison, exchange and verification. This also inspired and boosted the development of an assortment of format conversion tools and libraries (Chambers et al., 2012; Griss et al., 2012) and stand-alone software for visualization of the content of the files in standard formats (Ghali et al., 2013) but, most importantly for the purpose of this work, enabled the possibility for Proteopathogen to benefit from the mzIdentML adopted standard for identification results, incorporating it as the input data format and using it as inspiration for information display.

More recently, the most comprehensive, up to the current date, online *C. albicans* proteomics data repository was developed and integrated in PeptideAtlas (Vialas et al., 2013). These

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

publicly available *C. albicans* results have been used to establish a new version of Proteopathogen with a solid foundation.

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

Materials and methods

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

As described in Ref. (Vialas et al., 2013) the data sets come from a range of experiments including yeast to hypha transition assays, membrane protein extractions and a set of phosphoprotein enrichment approaches. In all cases, cells from the clinical isolates SC5314 were grown in YPD medium. For obtaining cells in hyphal form, either heat-inactivated fetal bovine serum or Lee medium pH 6.7 was used. As for the mass spectrometry, spectra were acquired in different set ups and platforms in a data-dependent manner. A summary of the experiments set ups and conditions is shown in Table 1.

Consistently with the PeptideAtlas project principles, the MS output files were processed through the Trans Proteomic Pipeline. The steps involved, first, sequence database searching using X! Tandem with k-score (MacLean et al., 2006) and a custom sequence database obtained from Candida Genome Database (Costanzo et al., 2006) with appended decoy counterparts and common contaminants for peptide-to-spectrum matching and FDR assessment. Then the post-processing validation tools PeptideProphet (Choi and Nesvizhskii, 2008), ProteinProphet (Nesvizhskii et al., 2003) and iProphet (Shteynberg et al., 2011) provided filtered lists of peptides and proteins with high probabilities. And finally FDR was computed for different probability thresholds.

Each of the PeptideAtlas repository datasets consists on the MS output spectra files and a set of pepXML and protXML files with lists of high confidence peptide and proteins respectively. These were combined, independently for each dataset, by means of a custom script written in the Ruby scripting language (available in supplemental data) to create mzIdentML files (mzIdentML version 1.1.0) with the merged information. In order to check the files were generated correctly and ensure data quality they were all validated (semantic and MIAPEx-compliant validation) with mzidValidator (Ghali et al., 2013).

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

| Type of dataset | Number of MS output files | Instrument | PeptideAtlas datasets |
|---|---------------------------|--------------------------------|---|
| <i>Candida albicans</i> culture with SILAC labelling, digested protein extracts enriched in phosphopeptides IMAC/TiO2 | 57 | Orbitrap XL, Orbitrap Velos | PAe001976, PAe001977, PAe001978, PAe001979, PAe001980, PAe001984, PAe001985, PAe001986, PAe001987, PAe001988, PAe001989 |
| <i>Candida albicans</i> total protein extract, 2 Triple-TOF runs, 2ug and 4ug | 2 | Triple-TOF | PAe001983 |
| Hyphal form and yeast form total protein extracts | 8 | Orbitrap Velos | PAe002110, PAe002111 |
| LTQ membrane proteins (Cabezón et al., 2009) | 3 | LTQ | PAe001981 |
| LTQ proteins from acidic sub-proteome (Monteoliva et al., 2011) | 8 | LTQ | PAe001982 |

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

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

Results and discussion

A total number of sixteen mzIdentML files, corresponding to each of the PeptideAtlas repository datasets, grouped into five different experiments were compiled and used to initially populate the Proteopathogen database. These account for approximately 22,000 distinct peptides and 2600 different proteins that can be queried and viewed through the web interface.

Precisely, a stringent FDR cut-off at the PSM level set at 0.005, yields 21,883 peptides with 0.0024 FDR (peptide level) and 2577 proteins with 0.0170 FDR (protein level) as computed with Mayu, a software specifically designed to estimate accurate protein level error rates in large datasets (Reiter et al., 2009) (see supplemental Table 1).

The mzIdentML contents can be browsed for each file in Proteopathogen in a means inspi-

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

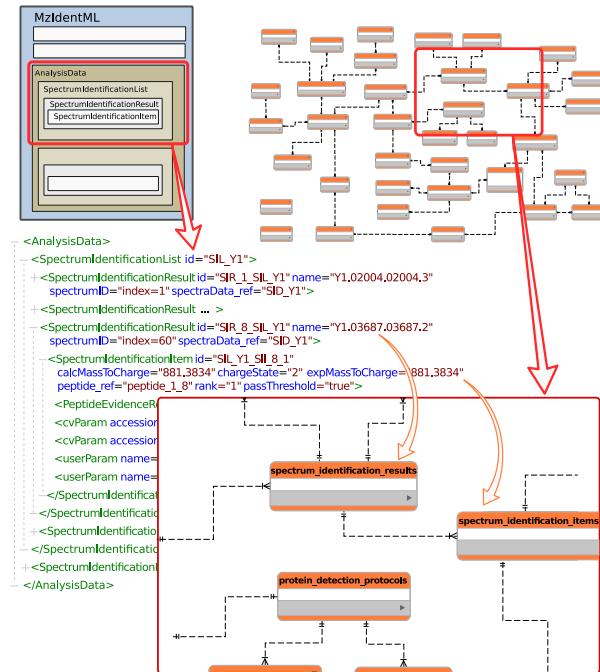


Figure 1. mzIdentML to database mapping. The MySQL schema was specifically designed to accommodate elements from the mzIdentML format. Figure shows the one-to-many relationship between the *<SpectrumIdentificationResult>* and *<SpectrumIdentificationItem>*elements.

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

Notably, the information Proteopathogen displays will depend on how complete the original mzIdentML files are. For instance, for files including the optional *<Fragmentation>* element under *<SpectrumIdentificationItem>*, Proteopathogen will display an annotated and interactive MS/MS spectrum. In addition, the optional *<cvParam>* and *<userParam>* elements, that descri-

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

be and annotate with controlled vocabularies and user-defined information respectively different elements throughout the file, might be more or less profuse depending on the software that created them.

EXPERIMENT short label: Candida albicans total protein extract, 2 TripleTOF runs, 2 and 4 ug
Protocol: view protocol description
Organism: Candida albicans (strain SC5314 / ATCC MYA-2876)
Contact: mihnae@farm.ucm.es
Date: 10-2011

.mzid File: C_albicans_total_protein_extract_TripleTOF.mzid

SPECTRUM IDENTIFICATION
Spectrum Identification: SI_Y1 - Spectrum Identification List: SIL_Y1

Spectrum Identification Results

| Spectrum Result ID | Spectrum Name | Spectrum ID |
|-----------------------|--------------------|-------------|
| SIR_59_SI_Candida2ug | Candida2ug,20727,2 | Index=1426 |
| SIR_60_SI_Candida_2ug | Candida2ug,20747,2 | index=1428 |
| ... | ... | ... |

Spectrum Identification Items

| SII ID | calc:m/z | exp:m/z | rank | charge | pass thrshld |
|--------------------------|----------|----------|------|--------|--------------|
| SIL_Candida_2ug_SII_60_1 | 617.3108 | 617.3113 | 1 | 2 | true |

Peptide Evidence(s)

| start | end | pre | post | is_decoy | DB sequence |
|-------|-----|-----|------|----------|-------------|
| 164 | 174 | G | A | false | orf19.1779 |

VDCDQVSAVLK
UNIMOD:4 [Carbamidomethyl] C 3 - avg mass delta: 57.0215

SPECTRUM IDENTIFICATION ITEM
*MS/MS spectrum is shown if <Fragmentation> element is present in the mzid file

Spectrum Identification Item PSI-MS CV Parameters

| | | |
|------------|---------------------|-------|
| MS:1001331 | x tandem:hyperscore | 322 |
| MS:1001330 | x tandem:expect | 0.024 |

Spectrum Identification Item User Parameters

| | |
|-----------------------------|----------|
| peptide_prophet_probability | 0.99962 |
| interpretor_probability | 0.994339 |

PROTEIN DETECTION

Protein Detection List

| | |
|-------------------------------|----------|
| ProteinDetection Id | PDL_1 |
| Protein Detection Protocol ID | PD_P_1 |
| Analysis Software | XITandem |

Protein Ambiguity Groups

| Protein Group ID | Accession(s) / Id(s) | Gene Name(s) |
|------------------|----------------------|--------------|
| PAG_51 | orf19.1770 | CYC1 |
| PAG_52 | orf19.1779 | MP65 |
| ... | ... | ... |

Protein Detection Hypothesis/Hypotheses

| PDI ID | Protein Description | pass thrshld | name | CGD ID |
|----------|---|--------------|------------|--------|
| PDH_52_3 | MP65 CGID:CAL0000936... Verified ORF; Cell surface mannoprotein; | true | orf19.1779 | CGD |

Protein Detection Hypothesis PSI-MS CV Parameters

| | | |
|------------|---------------------|-----|
| MS:1001093 | x tandem:hyperscore | 322 |
|------------|---------------------|-----|

Protein Detection Hypothesis User Parameters

| | |
|-------------------|------|
| n_inlist_proteins | 1 |
| probability | 1000 |
| ... | ... |

orf19.1779
HLFKSFVFTVANALAAPLAHIOHQIKEEKRAVHVTTTIVVVTGQDTTTFAAPPSVAESSSVSVNTEPPONHPTTQDVASAT
YPSSTDGSASSSSAASSSSQNSGSEPSGVGGSGAKG ITYSPYSDNGCKSE50IASEIAQLSGFDVIR LYV VDCDQVSAVLK
AKTSQSKIFAGIFDVNSIAGSIESLAEAVAGSGMD01TYVSIGELNVAGSATPSQIAYDEGRKALKAGITGPVIVDF, ...

Peptide Spectrum Matches

| Peptide Sequence | Spectrum Identification Item(s) | PSM count |
|--------------------|---------------------------------|-----------|
| ITYSPYSDNGCK | SIL_Candida4ug_SII_2757_1 | 1 |
| VDCDQVSAVLK | SIL_Candida2ug_SII_60_1 | 1 |
| ESQIASEIAQLSGFDVIR | SIL_Candida4ug_SII_4141_1 | 1 |

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

In addition to browsing through the contents of the stored mzIdentML files, Proteopathogen implements a query system yielding global results. That is, for a specific queried protein name, as found in the <ProteinDetectionHypothesis>name attribute, the search results display all the distinct peptide sequences found mapped into the protein sequence, regardless of the

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

experiment in which they were identified, and the supporting spectra for each peptide sequence, while keeping track of the original <SpectrumIdentification> and mzIdentML file. A peptide sequence may also be searched, obtaining, when found, the corresponding protein, or group of proteins, and the set of supporting spectra, again in global scope.

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

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

Conclusions

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

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

References

- Cabezón, V., Llama-Palacios, A., Nombela, C., Monteoliva, L., and Gil, C. (2009), Analysis of *Candida albicans* plasma membrane proteome., *Proteomics*, 9(20), 4770-86.
- Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012), A cross-platform toolkit for mass spectrometry and proteomics., *Nature biotechnology*, 30(10), 918-20
- Choi, H. and Nesvizhskii, A. I. (2008), Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics., *Journal of proteome research*, 7(1), 254-65.
- Costanzo, M. C., Arnaud, M. B., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., and Sherlock, G. (2006), The *Candida* Genome Database: facilitating research on *Candida albicans* molecular biology., *FEMS yeast research*, 6(5), 671-84.
- Deutsch, E. (2012), File formats commonly used in mass spectrometry proteomics, *Molecular & Cellular Proteomics*, 11(12), 1612-1621.
- Fernández-Arenas, E., Cabezón, V., Bermejo, C., Arroyo, J., Nombela, C., Diez-Orejas, R., and Gil, C. (2007), Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction., *Molecular & cellular proteomics : MCP*, 6(3), 460-478.
- Ghali, F., Krishna, R., Lukasse, P., Martínez-Bartolomé, S., Reisinger, F., Hermjakob, H., Vizcaíno, J. A., and Jones, A. R. (2013), Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML., *Molecular & cellular proteomics : MCP*, 12(11), 3026-35.
- Gil-Bona, A., Llama-Palacios, A., Parra, C. M., Vivanco, F., Nombela, C., Monteoliva, L., and Gil, C. (2014), Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in *Candida albicans*., *Journal of proteome research*, 14(1), 142-53.

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

- Gow, N. and van de Veerdonk, F. (2011), *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization, *Nature Reviews*, 10(2), 112-122.
- Griss, J., Reisinger, F., Hermjakob, H., and Vizcaíno, J. A. (2012), jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats., *Proteomics*, 12(6), 795-8.
- Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S. J., Selley, J. N., Searle, B. C., Shofstahl, J., Seymour, S. L., Julian, R., Binz, P.-A., Deutsch, E. W., Hermjakob, H., Reisinger, F., Griss, J., Vizcaíno, J. A., Chambers, M., Pizarro, A., and Creasy, D. (2012), The mzIdentML data standard for mass spectrometry-based proteomics results., *Molecular & cellular proteomics : MCP*, 11(7), M111.014381.
- Madeo, F., Herker, E., and Wissing, S. (2004), Apoptosis in yeast, *Current opinion in microbiology*, 7(6), 655-660
- MacLean, B., Eng, J., Beavis, R., and McIntosh, M. (2006), General framework for developing and evaluating database scoring algorithms using the TANDEM search engine, *Bioinformatics*, 22(22), 2830-2832.
- Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpf, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011), mzML-a community standard for mass spectrometry data., *Molecular & cellular proteomics : MCP*, 10(1), R110.000133.
- Martínez-Bartolomé, S., Binz, P.-A., and Albar, J. P. (2014), The Minimal Information about a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative., *Methods in molecular biology* (Clifton, N.J.), 1072, 765-80.
- Monteoliva, L., Martinez-Lopez, R., Pitarch, A., Hernaez, M. L., Serna, A., Nombela, C., Albar, J. P., and Gil, C. (2011), Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition, *Journal of Proteome Research*, 10(2), 502-517.
- Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003), A statistical model for identifying proteins by tandem mass spectrometry., *Analytical chemistry*, 75(17), 4646-58.
- Ramsdale, M. (2008), Programmed cell death in pathogenic fungi, *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1783(7), 1369-1380.

PROTEOPATHOGEN2, ADAPTED TO MZIDENTML

Reiter, L., Claassen, M., Schrimpf, S. S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009), Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry, Molecular & Cellular Proteomics, 8(11), 2405-2417.

Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I. (2011), iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates, Molecular & Cellular Proteomics, 10(12), M111.007690-M111.007690.

Vialás, V., Nogales-Cadenas, R., Nombela, C., Pascual-Montano, A., and Gil, C. (2009), Proteopathogen, a protein database for studying *Candida albicans*-host interaction., Proteomics, 9(20), 4664-8.

Vialás, V., Perumal, P., Gutierrez, D., Ximénez-Eembún, P., Nombela, C., Gil, C., and Chaffin, W. L. (2012), Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells., Proteomics, 12(14), 2331-2339.

Vialas, V., Sun, Z., Loureiro Y Penha, C. V., Carrascal, M., Abián, J., Monteoliva, L., Deutsch, E. W., Aebersold, R., Moritz, R. L., and Gil, C. (2013), A *Candida albicans* PeptideAtlas., Journal of proteomics, 97, 62-8.

CREACIÓN DE UN PEPTIDEATLAS DE *Candida albicans*

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

Richard Dawkins
Unweaving the rainbow

A Candida albicans PeptideAtlas

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

Journal of Proteomics 2014, 97, 62-68

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.

A *Candida albicans* PEPTIDEATLAS

Introduction

Candida albicans is a fungus of great clinical importance. In addition to asymptotically colonizing mucous membranes as a commensal in a large percentage of the population, it may cause severe opportunistic infections in specific cases such as patients with weakened immune defenses, a common circumstance in cancer and AIDS patients. *C. albicans* infections are also a threat to patients in post-surgical situations and intensive care unit stays. In this respect, invasive candidiasis remains nowadays one of the major types of nosocomial infections and a challenge in terms of economical and health costs (Wisplinghoff et al., 2004; Moran et al., 2010; Tong et al., 2008). 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 (Pitarch et al., 2006a,b; Fernández-Arenas et al., 2007; Pitarch et al., 2011).

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 (Vizcaíno et al., 2013) database includes only 15 experiments accounting for 1786 identified proteins. The more *C. albicans*-focused Proteopathogen database (Vialás et al., 2009) 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 (Craig and Beavis, 2004)) or Tranche (Smith et al., 2011) 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 (Costanzo et al., 2006), 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 (Desiere et al., 2006) 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 (Farrah et al., 2011) and urine), microbial builds (*Halobacterium* (Van et al., 2008), *Mycobacterium tuberculosis* (Schubert et al., 2013), *Streptococcus* (Lange et al., 2008), *Leptospira*, *Plasmodium* (Lindner et al., 2013), *Saccharomyces* (King et al., 2006) and *Schizosaccharomyces* (Gunaratne et al., 2013)); invertebrate builds (*Caenorhabditis elegans*, *Drosophila*

A *Candida albicans* PEPTIDEATLAS

(Loevenich et al., 2009) and *Apis mellifera* (Chan et al., 2011)); and a pig and a bovine milk (Bislev et al., 2012) builds. The PeptideAtlas project, as a multi-species compendium of proteomes, is continuously increasing its biological diversity. The recent *Schizosaccharomyces pombe* atlas (Gunaratne et al., 2013) 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) (Deutsch et al., 2008a) or SWATH-MS (Gillet et al., 2012) will enable a starting point for future targeted proteomics workflows in *C. albicans*.

Materials and methods

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 (Cabezón et al., 2009), morphological yeast to hypha transition experiments (Monteoliva et al., 2011) and phosphoprotein enrichment treatments. The combination of these diverse datasets resulted in an unprecedented overall coverage of the *C. albicans* proteome. Protein samples were obtained as previously described in (Monteoliva et al., 2011). 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.

A *Candida albicans* PEPTIDEATLAS

| #Exp | 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 | OrbitrapVelos | 3 |
| 4 | SILAC_phos_OrbitrapVelos_2 | SILAC. IMAC+TiO2 | OrbitrapVelos | 3 |
| 5 | SILAC_phos_OrbitrapVelos_3 | SILAC. IMAC+TiO2 | OrbitrapVelos | 3 |
| 6 | SILAC_phos_OrbitrapVelos_4 | SILAC. IMAC+TiO2 | OrbitrapVelos | 3 |
| 7 | SILAC_phos_OrbitrapXL_1A | SILAC. IMAC | OrbitrapXL | 11 |
| 8 | SILAC_phos_OrbitrapXL_1A_TiO2 | SILAC. IMAC+TiO2 | OrbitrapXL | 5 |
| 9 | SILAC_phos_OrbitrapXL_1B | SILAC. IMAC | OrbitrapXL | 6 |
| 10 | SILAC_phos_OrbitrapXL_1B_TiO2 | SILAC. IMAC+TiO2 | OrbitrapXL | 6 |
| 11 | SILAC_phos_OrbitrapXL_2 | SILAC. IMAC | OrbitrapXL | 6 |
| 12 | SILAC_phos_OrbitrapXL_3 | SILAC. IMAC | OrbitrapXL | 6 |
| 13 | SILAC_phos_OrbitrapXL_4 | SILAC. IMAC | OrbitrapXL | 5 |
| 14 | Calb_extract_3TOF | - | Triple TOF | 2 |
| 15 | Hyphal_extract_OrbitrapVelos | - | Orbitrap Velos | 4 |
| 16 | Yeast_extract_OrbitrapVelos | - | Orbitrap Velos | 4 |

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

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) (Deutsch et al., 2010), 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 (Martens et al., 2011), searched using X!Tandem (Craig and Beavis, 2004) with the K-score algorithm plug-in (MacLean et al., 2006) and the output search results were converted to the search engine-independent pepXML format (Keller et al., 2005).

The target fasta sequence file used for the search was obtained from the Candida Genome Database (CGD) (Costanzo et al., 2006) (Assembly 21)

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 (Keller et al., 2002) was then run on the search results to model the distributions of correctly and incorrectly assigned Peptide-to-Spectrum Matches (PSMs). It then

A *Candida albicans* PEPTIDEATLAS

assigns probabilities of being correct for each PSM, yielding a sensitive and flexible approach to report results in a comparable manner. Next, iProphet (Shteynberg et al., 2011) 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 (Nesvizhskii et al., 2003) 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 (Reiter et al., 2009) 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 (Farrah et al., 2011) and successfully applied to other builds such as the bovine milk and mammary gland PeptideAtlas (Bislev et al., 2012).

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 the generic feature file *C. albicans* SC5314 versionA21 s02m05r10 features.gff obtained from CGD.

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.

Results and discussion

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* PeptideAtlas.

The PSM assignment and protein inference processes were conducted by means of the

A *Candida albicans* PEPTIDEATLAS

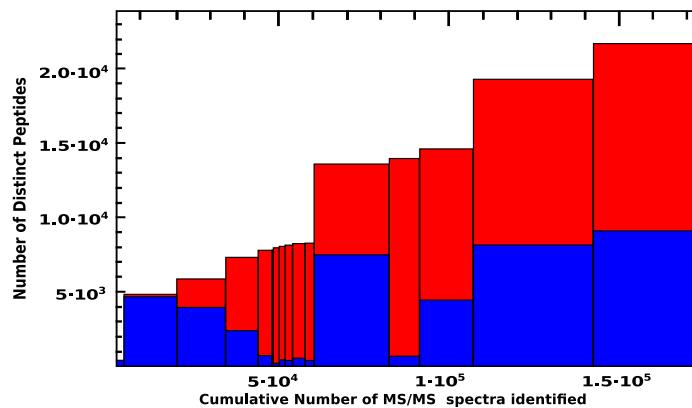


Figure 1. Histogram showing the cumulative number of distinct peptides in the *C. albicans* PeptideAtlas. Each bar represents a different experiment that has contributed to the build. Bar width is proportional to the number of high confidence PSMs. Height of the blue section of the bar represents the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

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 (Tabas-Madrid et al., 2012). 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

A *Candida albicans* PEPTIDEATLAS

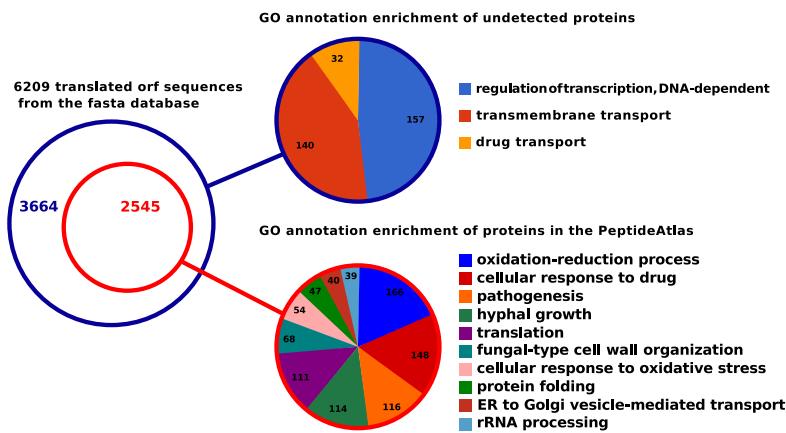


Figure 2. Gene Ontology annotation enrichment analysis for both the covered and undetected proteome subsets. All shown GO annotations correspond to the biological process ontology and were found significant for a p-value cut-off below 0.01.

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 (Gunaratne et al., 2013). 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 (Gunaratne et al., 2013; Ding et al., 2013; Simicevic et al., 2013). 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.

A *Candida albicans* PEPTIDEATLAS

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).

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) (Vialás et al., 2012). Proteins 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 (Saville et al., 2003).

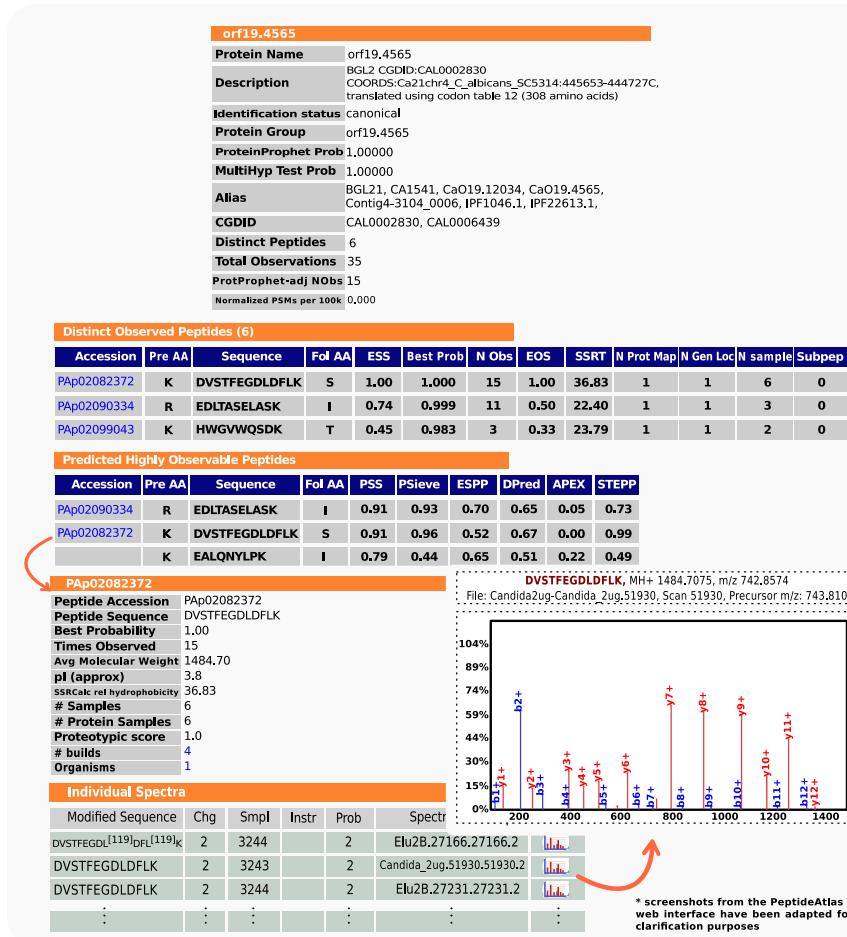
Within these groups, a selected set of proteins of interest present in the atlas, are the adhesins from the ALS family with a role in invasiveness Als2p and Als3p; those required for cell wall biogenesis and organization glycosidases Phr1p, Phr2p and Utr2p; mannosyltransferases Pmt1p, Pmt4 and Pmt6; 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 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 (Mallick et al., 2007).

A *Candida albicans* PEPTIDEATLAS



A *Candida albicans* PEPTIDEATLAS

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.

Conclusions

This *C. albicans* PeptideAtlas build provides empirical identification evidence for 21,938 unique peptides including 421 phosphopeptides at a 0.24 % peptide-level FDR that account for a high-confidence set (as defined in [14]) of 2562 canonical proteins at a 1.2% protein-level FDR representing thus a significant advance in the proteomic characterization of *C. albicans*. Through the web interface, an important set of tools are made available to the scientific community, enabling a solid foundation to study different basic biological processes like dimorphism, signal transduction, apoptosis and the interaction with the human host. Furthermore, its value as a resource for proteotypic peptide selection is of great potential interest for future SRM experiments. The current version of the PeptideAtlas can be found at: https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=323 and the version including PTM results at: https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=324

References

- Bislev, S., Deutsch, E., and Sun, Z. (2012), A Bovine PeptideAtlas of milk and mammary gland proteomes, *Molecular & Cellular Proteomics*, 12(18), 2895-2899.
- Cabezón, V., Llama-Palacios, A., Nombela, C., Monteoliva, L., and Gil, C. (2009), Analysis of *Candida albicans* plasma membrane proteome., *Proteomics*, 9(20), 4770-86.
- Chan, Q. W. T., Parker, R., Sun, Z., Deutsch, E. W., and Foster, L. J. (2011), A honey bee (*Apis mellifera* L.) PeptideAtlas crossing castes and tissues., *BMC genomics*, 12(1), 290.
- Costanzo, M. C., Arnaud, M. B., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., and Sherlock, G. (2006), The *Candida* Genome Database: facilitating research on *Candida albicans* molecular biology., *FEMS yeast research*, 6(5), 671-84.
- Craig, R. and Beavis, R. C. (2004), TANDEM: matching proteins with tandem mass spectra., *Bioinformatics*, 20(9), 1466-7.

A *Candida albicans* PEPTIDEATLAS

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006), The PeptideAtlas project., Nucleic acids research, 34(Database issue), D655-8.

Deutsch, E., Lam, H., and Aebersold, R. (2008), Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics, Physiological genomics, 33(1), 18-25.

Deutsch, E. E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., and Aebersold, R. (2010), A guided tour of the TransProteomic Pipeline, Proteomics, 10(6), 1150-1159.

Ding, C., Chan, D. W., Liu, W., Liu, M., Li, D., Song, L., Li, C., Jin, J., Malovannaya, A., Jung, S. Y., Zhen, B., Wang, Y., and Qin, J. (2013), Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements, Proceedings of the National Academy of Sciences of the United States of America, 110(17), 6771-6.

Farrah, T., Deutsch, E. W., Omenn, G. S., Campbell, D. S., Sun, Z., Bletz, J. a., Mallick, P., Katz, J. E., Malmström, J., Ossola, R., Watts, J. D., Lin, B., Zhang, H., Moritz, R. L., and Aebersold, R. (2011), A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas., Molecular & Cellular Proteomics, 10(9), M110.006353.

Fernández-Arenas, E., Cabezón, V., Bermejo, C., Arroyo, J., Nombela, C., Diez-Orejas, R., and Gil, C. (2007), Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction., Molecular & cellular proteomics: MCP, 6(3), 460-478.

Gillet, L. C., Navarro, P., Tate, S., Räst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012), Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis., Molecular & Cellular Proteomics, 11(6), O111.016717.

Gunaratne, J., Schmidt, A., Quandt, A., Neo, S. P., Sarac, O. S., Gracia, T., Loguercio, S., Ahrne, E., Li Hai Xia, R., Tan, K. H., Loessner, C., Bahler, J., Beyer, A., Blackstock, W., and Aebersold, R. (2013), Extensive Mass Spectrometry-Based Analysis of the Fission Yeast Proteome: The *S. pombe* PeptideAtlas, Molecular & Cellular Proteomics, pp. 1741-1751.

A *Candida albicans* PEPTIDEATLAS

Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002), Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search., *Analytical chemistry*, 74(20), 5383-92.

Keller, A., Eng, J., Zhang, N., Li, X.-j., and Aebersold, R. (2005), A uniform proteomics MS/MS analysis platform utilizing open XML file formats, *Molecular systems biology*, 1(August 2005), 2005.0017.

King, N. L., Deutsch, E. W., Ranish, J. A., Nesvizhskii, A. I., Eddes, J. S., Mallick, P., Eng, J., Desiere, F., Flory, M., Martin, D. B., Kim, B., Lee, H., Raught, B., and Aebersold, R. (2006), Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas., *Genome biology*, 7(11), R106

Lindner, S. E., Swearingen, K. E., Harupa, A., Vaughan, A. M., Sinnis, P., Moritz, R. L., and Kappe, S. H. I. (2013), Total and putative surface proteomics of malaria parasite salivary gland sporozoites., *Molecular & Cellular Proteomics*, 12(5), 1127-43.

Loevenich, S. N., Brunner, E., King, N. L., Deutsch, E. W., Stein, S. E., Aebersold, R., and Hafen, E. (2009), The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation., *BMC bioinformatics*, 10, 59.

MacLean, B., Eng, J., Beavis, R., and McIntosh, M. (2006), General framework for developing and evaluating database scoring algorithms using the TANDEM search engine, *Bioinformatics*, 22(22), 2830-2832.

Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold, R. (2007), Computational prediction of proteotypic peptides for quantitative proteomics., *Nature biotechnology*, 25(1), 125-31.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Rompp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011), mzML-a community standard for mass spectrometry data., *Molecular & cellular proteomics : MCP*, 10(1), R110.000133.

Monteoliva, L., Martinez-Lopez, R., Pitarch, A., Hernaez, M. L., Serna, A., Nombela, C., Albar, J. P., and Gil, C. (2011), Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition, *Journal of Proteome Research*, 10(2), 502-517.

A *Candida albicans* PEPTIDEATLAS

- Moran, C., Grussemeyer, C. A., Spalding, J. R., Benjamin, D. K., and Reed, S. D. (2010), Comparison of costs, length of stay, and mortality associated with *Candida glabrata* and *Candida albicans* bloodstream infections., *American journal of infection control*, 38(1), 78- 80.
- Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003), A statistical model for identifying proteins by tandem mass spectrometry., *Analytical chemistry*, 75(17), 4646-58.
- Pitarch, A., Nombela, C., and Gil, C. (2006a), *Candida albicans* biology and pathogenicity: insights from proteomics., *Methods of biochemical analysis*, 49, 285-330.
- Reiter, L., Rinner, O., Picotti, P., HÄttenhain, R., Beck, M., Brusniak, M.-Y., Hengartner, M. O., and Aebersold, R. (2011), mProphet: automated data processing and statistical validation for large-scale SRM experiments., *Nature methods*, 8(5), 430-5.
- Saville, S. P., Lazzell, A. L., Monteagudo, C., and Lopez-Ribot, J. L. (2003), Engineered control of cell morphology *in vivo* reveals distinct roles for yeast and filamentous forms of *Candida albicans* during infection., *Eukaryotic cell*, 2(5), 1053-60.
- Schubert, O. T., Mouritsen, J., Ludwig, C., RÄst, H. L., Rosenberger, G., Arthur, P. K., Claassen, M., Campbell, D. S., Sun, Z., Farrah, T., Gengenbacher, M., Maiolica, A., Kaufmann, S. H. E., Moritz, R. L., and Aebersold, R. (2013), The Mtb Proteome Library: A Resource of Assays to Quantify the Complete Proteome of *Mycobacterium tuberculosis*., *Cell host & microbe*, 13(5), 602-12.
- Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, a. I. (2011), iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates, *Molecular & Cellular Proteomics*, 10(12), M111.007690-M111.007690.
- Simicevic, J., Schmid, A. W., Gilardoni, P. A., Zoller, B., Raghav, S. K., Krier, I., Gubelmann, C., Lisacek, F., Naef, F., Moniatte, M., and Deplancke, B. (2013), Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics., *Nature methods*, advance on.
- Smith, B. E., Hill, J. A., Gjukich, M. A., and Andrews, P. C. (2011), Tranche distributed repository and ProteomeCommons.org., *Methods in molecular biology* (Clifton, N.J.), 696, 123-45.

A *Candida albicans* PEPTIDEATLAS

Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012), GeneCo-dis3: a non-redundant and modular enrichment analysis tool for functional genomics., Nucleic acids research, 40(Web Server issue), W478-83.

Tong, K. B., Murtagh, K. N., Lau, C., and Seifeldin, R. (2008), The impact of esophageal candidiasis on hospital charges and costs across patient subgroups., Current medical research and opinion, 24(1), 167-74.

Van, P. T., Schmid, A. K., King, N. L., Kaur, A., Pan, M., Whitehead, K., Koide, T., Facciotti, M. T., Goo, Y. A., Deutsch, E. W., Reiss, D. J., Mallick, P., and Baliga, N. S. (2008), Halobacterium salinarum NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage., Journal of proteome research, 7(9), 3755-64.

Vialás, V., Nogales-Cadenas, R., Nombela, C., Pascual-Montano, A., and Gil, C. (2009), Proteopathogen, a protein database for studying *Candida albicans*-host interaction., Proteomics, 9(20), 4664-8.

Vialás, V., Perumal, P., Gutierrez, D., Ximénez-Erbún, P., Nombela, C., Gil, C., and Chaffin, W. L. (2012), Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells., Proteomics, 12(14), 2331-2339.

Vizcaíno, J. A., Coté, R. G., Csordas, A., Dianes, J. a., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O Kelly, G., Schoenegger, A., Ovelleiro, D., Pérez-Riverol, Y., Reisinger, F., RÁos, D., Wang, R., and Hermjakob, H. (2013), The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013., Nucleic acids research, 41(Database issue), D1063-9.

Wisplinghoff, H., Bischoff, T., Tallent, S. M., Seifert, H., Wenzel, R. P., and Edmond, M. B. (2004), Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study., Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 39(3), 309-17.

**A comprehensive *Candida albicans* PeptideAtlas build enables
deep proteome coverage**

Abstract

To provide new and expanded proteome documentation of the opportunistically pathogen *Candida albicans*, we have developed new protein extraction and analysis routines to provide a new, extended and enhanced version of the *C. albicans* PeptideAtlas. Two new datasets, resulting from experiments consisting of exhaustive subcellular fractionations and different growing conditions, plus two additional datasets from previous experiments on the surface and the secreted proteomes, have been incorporated to increase the coverage of the proteome. High resolution precursor mass spectrometry (MS) and ion trap tandem MS spectra were analyzed with three different search engines using a database containing allele-specific sequences. This novel approach combined with the post-processing and filtering implemented in the Trans Proteomic Pipeline consistently used in the PeptideAtlas project resulted in 49372 additional peptides and 1630 more proteins identified in the new *C. albicans* PeptideAtlas with respect to the previous build. A total of 71310 peptides and 4174 canonical (minimal non-redundant set) proteins (4115 if one protein per pair of alleles is considered) were identified representing 66 % of the 6218 proteins in the predicted proteome. This makes the new PeptideAtlas build the most comprehensive *C. albicans* proteomics resource available and the only large-scale one with detections of individual alleles.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

Introduction

Candida albicans, an inhabitant fungus of the gastrointestinal and genitourinary tracts in the human microbiota, may under certain circumstances (e.g., present in patients with a weakened immune system from AIDS or patients in an intensive care unit), become opportunistically pathogenic and hence become the etiological agent of a severe type of infection called candidiasis. In the search for diagnostic and prognostic biomarkers for candidiasis, a large assortment of proteomics studies have been performed (Pitarch et al., 2011), or with the objective to better understand clinically relevant biological processes such as the interaction with cells of the immune system (Fernández-Arenas et al., 2007; Cheng et al., 2012; Gow and van de Veerdonk, 2011), apoptosis (Ramsdale, 2008; Hao et al., 2013) or the virulence-associated morphological yeast-to-hypha switch (Monteoliva et al., 2011; Vialás et al., 2012).

However, despite the extensive efforts on clinical aspects from the proteomics view (Pitarch et al., 2006a; Cabezón et al., 2009; Pitarch et al., 2009; Rupp, 2004), online public proteomic repositories were sparse. Our previously published *C. albicans* PeptideAtlas (Vialás et al., 2013) described the first large-scale public proteomic resource for the study of this opportunistic pathogenic fungus. With over 2500 proteins and representing approximately 41 % of the predicted proteome, the *C. albicans* PeptideAtlas attained unprecedented proteome coverage and is still is the first human fungal pathogen present in the PeptideAtlas project (Desiere et al., 2006; Deutsch et al., 2015). However, this coverage lags far behind the coverage for other species, such as the yeast species *Saccharomyces cerevisiae* (61 %) (King et al., 2006) and *Schizosaccharomyces pombe* (71 %) (Gunaratne et al., 2013). To bridge this gap we have added additional high resolution precursor MS datasets based on purification strategies that collectively strive for maximizing the coverage of the detectable proteome. One of the approaches consists of an exhaustive subcellular fractionation based on differential sequential centrifugations followed by protein-level separation on SDS-PAGE; the other approach is based on a peptide-level separation by OFFGEL preparative isoelectric focusing of peptides (Ros et al., 2002; Hörrt et al., 2006) from a pool from five different culture conditions. In addition, two high-resolution *C. albicans* MS datasets corresponding to published works on the secreted proteome (Gil-Bona et al., 2014) and on the yeast and hyphal cell forms surface proteome (Gil-Bona et al., 2015) have also been included in the compilation of MS data files. These new datasets, along with the previously stored raw MS datafiles comprising the former version of the *C. albicans* PeptideAtlas (Vialás et al., 2013), have all been reprocessed and analyzed through the Trans Proteomics Pipeline, TPP (Keller et al., 2005; Deutsch et al., 2015) using multi-search search engine approach utilizing a sequence database with allele-specific sequences to generate the most comprehensive existing *C. albicans* proteomics resource with unprecedented proteome coverage.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

| Sample (as named in the web interface) | Experiment | PASS id | MS type | Fractions/Replicates | # spectra files |
|--|------------|--|--------------------|---|-----------------|
| Calb_subcel_fract | PASS00402 | Differential sequential centrifugations and SDS-PAGE | LTQ XL Orbitrap | FractionA x20 FractionB x20 FractionC 2x20 FractionD x20 | 100 |
| Calb_offGel | PASS00447 | OFFGEL Preparative Isoelectric focusing of peptides | LTQ XL Orbitrap | 24 OFFGEL fractions | 24 |
| Calb_ves_secreteome | PASS00408 | (Gil-Bona et al., 2014) | LTQ Orbitrap Velos | Wt and strains Vesicle-free: x3 Vesicles: x3 | 12 |
| Calb_surfome | PASS00446 | (Gil-Bona 2015) | LTQ Orbitrap Velos | Yest form: x3 serum-induced hyp-hae: x3 I.serum-induced hyp-hae: x2 Lee-induced hyphae: x4 | 14 |

Table 1. Overview of the new datasets that were reprocessed together with the datasets from the previous version of the PeptideAtlas to produce the new build.

Experimental procedures

Cell strains and culture media

The *C. albicans* strain used throughout this development is the widely adopted wild-type SC5314, the strain used as the reference sequence in Candida Genome Database, CGD (Costanzo et al., 2006), which was also used as the reference sequence database for spectra to peptide assignment. The exhaustive subcellular fractionation is derived from protoplasts obtained from cells cultivated in YED culture medium (1% D-glucose, 1% Difco yeast extract and 2% agar). For fractionation based on preparative isoelectric focusing of tryptic peptides, the following culture conditions were used: YED medium + 5% H₂O₂ at 30° C; YED medium at 42° C; YED medium at 30° C until stationary phase; RPMI 1640 medium (supplemented with 5% Fetal Bovine Serum, FBS) at 37° C and Minimal Medium at 30° C.

Cell lysis and protein digestion

C. albicans cells from the different culture conditions were washed three times with ice-cold PBS and then scraped and collected by centrifugation at 5000g. Pellets were resuspended in lysis buffer (30mM Tris-HCl pH 8.5, 7M urea, 2M thiourea, 4% CHAPS, 1% protease inhibitor cocktail -Roche- and 0.5% PMSF). An equal volume of 0.4-0.6 mm diameter glass beads

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

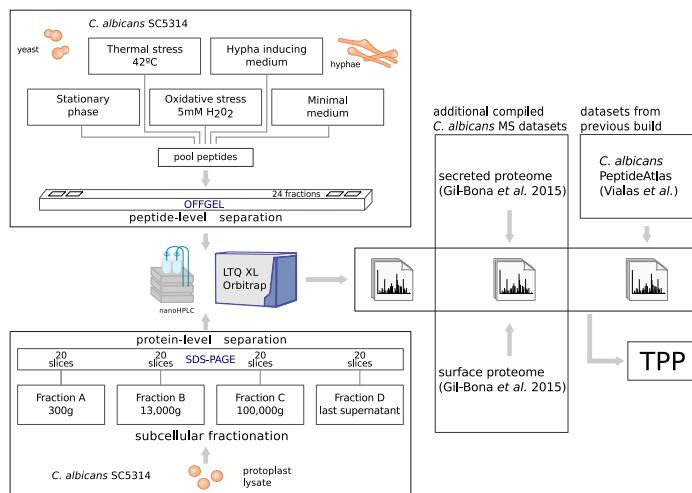


Figure 1. The strategies implemented *ad hoc* to maximize the coverage of the proteome consist of a subcellular fractionation based on increasing centrifugation speeds followed by exhaustive protein- level separation on SDS-PAGE; and OFFGEL peptide-level separation from a pool of peptides from different growing and culture conditions. The MS output results, combined with additional MS datasets from works on the secreted and surface proteomes, along with the datasets in the first atlas were all processed with the TPP.

was added. Subsequently, cells were disrupted in a FastPrep cell breaker. Supernatants were separated by centrifugation at 3000g for 10 min and protein quantitation was measured using a Bradford assay (Biorad). Equal amounts of each condition (250 µg/sample) were pooled and denatured by adding 25 mM DTT for 30 min at 60°C. Then, samples were loaded into an Amicon (Nanosep 10K Omega; Pall Corporation) and centrifuged 45 min at 12000 g. Samples were washed twice with DB2 buffer (20mM TEAB, 0,5 % sodium deoxycolate) and alkylated with 50 mM Iodoacetamide during 20 min in the dark. After twice washes with DB2, digestion was performed by adding sequence grade-modified trypsin (Roche) at an enzyme to substrate ratio of 1:50. After 12h in the dark at RT, peptides were collected into a clean collection tube and the Amicon was washed with DB2 and the flow-through was collected with samples acidified with 0.5 % (v/v) TFA. Any protein precipitation was separated by centrifugation for 5 min at 16000 g.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

OFFGEL peptide fractionation

For peptide isoelectric focusing (IEF) separation, the resulting peptide mixture (1.2mg in total) was resuspended in a buffer containing 6 % glycerol and 1.2 % ampholytes in the 3-10 pH linear OFFGEL buffer (7M Urea, 2M thiourea, 1 % DTT) (GE Healthcare, Uppsala, Sweden). Sample volumes of 150 μ l were loaded onto a commercially available 24-cm IPG strip with a linear 3-10 pH gradient (GE Healthcare) after rehydration of the gel for 20 min in a well of 40 μ l rehydration solution. Cover fluid (mineral oil, Agilent Technologies) was applied to both ends of the gel strip. Electrofocusing of the peptides was performed at 20C and 50 μ A until 50 kWh were reached using an Agilent 3100 OFFGEL fractionator (Agilent Technologies) following the manufacturer instructions. Fractions were recovered, peptides extracted from each well with 2 % TFA and desalting by passing through a home-made column packed with Poros 50 R2 resin (Applied Biosystem, Foster City, CA, USA). Peptides were eluted with 50 % ACN in 0.1 % TFA and the fractions were dried and reconstituted in 0.1 % formic acid just before LC-MS analysis.

Subcellular fractionation

Sequential incremental centrifugations was used to selectively enrich for different types of membranes and organelles in the pellets, while also collecting soluble proteins from the supernatant (Figure 1). First, protoplasts were obtained as described in (Pitarch et al., 2006a) and then lysed using a combination of osmotic shock and Dounce homogenization. Protoplast lysates were then subjected to a low centrifugal force, 300 g, resulting in a pellet, P 300, containing unlysed cells and large debris (Fraction A), and a supernatant, S 300, that is subsequently, subjected to increasing centrifugation speeds of 13,000 and 100,000 g. These steps, respectively, generate a pellet, P 13000, containing vacuolar and plasma membrane and other structures such as ER, mitochondria and nuclei (Fraction B), and a pellet, P 100000, containing Golgi membranes and transport vesicles (Fraction C). The supernatant of the last centrifugation containing soluble proteins was also collected, S 100000 (Fraction D). For each of these 4 fractions (P 300 , P 13000 , P 100000 and S 100000), 120 μ g of protein was separated by one- dimensional SDS-PAGE 4-20 % Bis-Tris gels (mini-protean TGX Stain-free precast Gels, BioRad). The gel was stained with Coomassie blue and each lane was cut into 20 bands. Gel slices were cut into 1mm 3 cubes, washed twice with water, dehydrated with 100 % acetonitrile (ACN), and incubated with 10 mM DTT in 50 mM NH₄HCO₃ for 30 min at 56°C for protein reduction. The resulting solution was subsequently alkylated by incubation with 55 mM iodoacetamide in 50 mM NH₄HCO₃ for 20 min at room temperature in the dark. The gel pieces were washed with 50 % ACN, and then washed again with 10 mM NH₄HCO₃, dehydrated with 100 % ACN, and then dried in a vacuum concentrator. The gel pieces were rehydrated by adding sequence grade-modified trypsin (Roche) 1:20 in 50 mM NH₄HCO₃ and

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

incubated overnight at room temperature in the dark for protein digestion. Supernatants were transferred to clean tubes, and gel pieces were incubated in 50 mM NH₄HCO₃ at 50°C for 1 h. The supernatants were collected and the remaining peptides were extracted by incubation with 5% formic acid for 15 min and with 100 % ACN for 15 min more. The extracts were combined, and the organic solvent was removed in a vacuum concentrator.

Compilation of additional *C. albicans* MS datasets

In addition to the current datasets, two high-resolution precursor *C. albicans* MS datasets were compiled in order to contribute to the new PeptideAtlas build, extending the coverage in two more specific niches, the set of secreted proteins obtained following the method described in (Gil-Bona et al., 2014) and the set of surface-exposed proteins, also termed surfacome, of both hyphal and yeast form cells as described in (Gil-Bona et al., 2015).

LC-MS/MS

All the samples obtained in the exhaustive subcellular fractionation and the OFFGEL peptide separation were analyzed in an LTQ XL Orbitrap (ThermoFisher) equipped with a nanoESI ion source. Samples were loaded into a chromatographic system consisting in a C18 pre-concentration cartridge (Agilent Technologies) connected to a 60 cm long, 100 µm i.d. C18 column (NanoSeparations) for the OFFGEL samples and a 15 cm long, 100 µm i.d. C18 column (Nikkyo Technos Co.) for the subcellular fractionation samples. For samples obtained using the OFFGEL approach, the injected volume was 8 % of the volume from each fraction whereas in the subcellular fractionation, one-third of the volume of each digested gel band was injected. High-resolution LC separation was performed at 0.25 µL/min using a 360-min acetonitrile gradient (OFFGEL samples) and at 0.4 µL/min in a 90-min acetonitrile gradient (subcellular fractionation samples). Both gradients ranged from 3 to 40 % (solvent A: 0.1 % formic acid, solvent B: acetonitrile 0.1 % formic acid). The HPLC system was composed of an Agilent 1200 capillary nano pump, a binary pump, a thermostated micro injector and a micro switch valve. The LTQ XL Orbitrap was operated in the positive ion mode with a spray voltage of 1.8 kV. The spectrometric analysis was performed in a data dependent mode, acquiring a full scan followed by 10 MS/MS scans (CID, collision energy 35 %) of the 10 most intense signals detected in the MS scan. The full MS (range 300-1800) was acquired in the Orbitrap with a resolution of 60.000. The MS/MS spectra were acquired in the linear ion trap. Precursor ion charge state screening was set up to select monoisotopic ions and reject singly charged ions. In all cases, dynamic exclusion was enabled with a repeat count of 1 and exclusion duration of 30 s.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

Post spectra acquisition processing

LC-MS/MS spectra files resulting from the subcellular fractionation and the OFFGEL approaches and the secreted and surface proteomes (Figure 1), in their native vendor-specific format, along with the meta data corresponding to each approach, were submitted to PeptideAtlas via the PeptideAtlas Submission System (PASS) on-line submission form with dataset identifications PASS00402, PASS00447, PASS00408 and PASS00446. LC-MS/MS spectra files were converted to XML-based HUPO-PSI-adopted standard format for mass spectrometry output, mzML (Martens et al., 2011). The protein sequence fasta file was obtained from the *Candida* Genome Database (C_albicans_SC5314_version_A22-s05-m01-r01). Unlike the previous *C. albicans* PeptideAtlas, for this new build the sequences in the database are allele-specific, taking advantage of the recent assembly of phased diploid *C. albicans* (Muzey et al., 2013). Sequences were appended with a set of common contaminant proteins from the cRAP (common Repository of Adventitious Proteins) set from the GPM (<http://www.thegpm.org/crap/>) and decoy counterparts for every entry to add up a total of 25168 entries.

Then database searches was performed using three different search engines: Comet (Eng et al., 2013), an open-source, freely available version of SEQUEST (Eng et al., 1994), X!Tandem (Craig and Beavis, 2004) with the k-score algorithm plugin (MacLean et al., 2006), and OMS-SA (Geer et al., 2004). The search parameters were established depending on the type of experiment and instrument (see supplementary table *database_search_parameters* for a list of parameters).

Following sequence database searching, we used the TPP tool suite to validate the results. First, PeptideProphet (Keller et al., 2002) creates a discriminant search engine-independent score, models distributions of correctly and incorrectly assigned peptide spectrum matches (PSMs) and computes PSM posterior probabilities. Next, iProphet (Shteynberg et al., 2011), was used to further refine the PSM-level probabilities and calculated distinct peptide-level probabilities using corroborating information from other PSMs in the dataset. ProteinProphet (Nesvizhskii et al., 2003) then was used to further refine peptide probabilities based on the Number of Sibling Peptides (NSP) that each peptide shares within a protein; it also groups and reports proteins with a protein-level probability estimated from peptide-level probabilities.

To assemble the *C. albicans* PeptideAtlas, all individual iProphet files from the 20 compiled datasets (16 corresponding to the previous build plus 4 new extensive datasets) were filtered at a variable probability threshold to reach a constant PSM-level FDR threshold of 0.001 across all datasets.

The new *Candida albicans* PeptideAtlas is made available at https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=443

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

The Mayu (Reiter et al., 2009) software designed for large-scale protein FDR estimation was used to report FDR values at different levels (PSM, unique peptides and protein-level) for the whole build based on a strategy that estimates the number of false positive protein identifications from the number of proteins containing false positive PSMs, including a correction for high proteome coverage.

Functional analysis and estimation of protein abundances

To perform functional analyses, we used the resources available at CGD GOSlim, a tailor made subset of GO terms specific to *Candida* biology, and GOSlimMapper, a software that maps a provided list of genes to the high-level set of GOSlim terms in either of the three ontologies. Protein abundances were estimated using the emPAI method (Ishihama et al., 2005) and the online software emPAI Calc (Shinoda et al., 2010) for the set of proteins identified in the subcellular fractionation approach since it represents the largest contribution to this PeptideAtlas build. The emPAI values were log transformed in order to obtain a normalized, symmetrical around 0 abundance scale that was apportioned in the group of the most abundant proteins (from the largest value up to the one representing percentile 0.85 in the scale); the group of proteins with high abundance (between 0.85 and 0.75 in the scale); proteins with abundance around the median (the two central quartiles); low abundant proteins (those with emPAI values between percentiles 0.25 and 0.15); and the least abundant proteins (those corresponding to percentile 0.15 and lower in the emPAI scale).

Results and discussion

Strategies for exhaustive proteome characterization

The experimental approaches were specifically designed to improve as much as possible of the *C. albicans* proteome, especially considering its great plasticity and variability. In the subcellular fractionation approach, (Figure 1) each of the fractions is enriched in different organelles and therefore ideally contributes with different subsets of proteins (Harford and Bonifacino, 2001). In Fraction A, the low centrifugal force enriches for larger membranes and complexes. The subsequent increasing centrifugation speed enriches Fraction B in other types of membranes (vacuolar, nuclear, endosomal and plasma) and also in Golgi complex and endoplasmic reticulum. Fraction C, the pellet resulting from the highest speed, contains some smaller structures like some endosomal membranes, parts of Golgi complex and transport vesicles. Finally, Fraction D, the final supernatant, contains soluble cytoplasmic and other released proteins. A total of 100 MS output files, corresponding to 20 slices from each of the four fractions run in SDS-PAGE (plus one extra replicate for fraction C) (Table 1) were obtained. This approach makes by itself the largest contribution to the *C. albicans* PeptideAtlas build

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

with over 650,000 spectra of which 350,499 could be identified, and allocated to 28,599 peptides (5,839 identified exclusively in this dataset) corresponding to roughly 3,000 proteins, 48% of the full proteome.

In the OFFGEL approach, the variability of the proteome was stimulated by the different growing conditions. The thermal and oxidative types of stress ideally enforce the cell to produce certain populations of proteins to face these culture conditions; the minimal medium makes cells adapt to deprivation of certain nutrients and therefore activate alternative mechanisms or pathways; hyphae generated in RPMI medium supplemented with FBS, provide a set of proteins inherent to this growing form; and finally, cells in the stationary phase also ideally contribute with proteins that would not be present in other more favorable conditions. These multiple growing conditions subjected to peptide-level separation by the OFFGEL system, generated 24 fractions and equivalent MS output files that make, as a dataset, the second largest contribution to the *C. albicans* PeptideAtlas with 460,000 spectra searched, 223,395 of them identified and assigned to 27,360 peptides (5,846 unique to this dataset) which, in turn, were assigned to more than 3,000 proteins. An overview of the contributions of each dataset to the entirety of the atlas is depicted in Figure 2.

Assessment of increment in proteome coverage

A total 229 MS runs (124 corresponding to the datasets implemented ad hoc, plus 26 corresponding to the additionally compiled datasets on secreted and surface proteomes, plus 79 datasets from the previous *C. albicans* PeptideAtlas build) generated 2,255,208 spectra of which more than one-third, 984,462, could be allocated to a peptide sequence. In the resulting outcome, for a PSM FDR threshold set at 0.10 %, 71,310 peptides are detected which can be explained by the minimal non-redundant set of 4174 canonical *C. albicans* protein sequences (4115 if only one protein sequence per pair of alleles is considered), representing 66 % of the 6218 (as of March 2015) predicted different protein sequences. With respect to the 22,000 peptides and 2545 proteins reported in analogous manner in the first version of the *C. albicans* PeptideAtlas, the multi-search engine reprocessing with the new LC-MS/MS datasets represents an increment of more than 3-fold in terms of peptides and 1.6-fold in the number of identified proteins.

One remarkable additional value in the *C. albicans* PeptideAtlas is the report of highly confident (ProteinProphet probability >0.9) identification of proteins corresponding to uncharacterized genes (following the terminology in CGD), i.e. those genes without previously known empirical evidence for a translated product. These amounted to 1564 in the previous PeptideAtlas build and has notably increased to 2860 (note that uncharacterized is unrelated to which of the alleles originates the protein product), representing an increment from one-third to almost two-thirds (63 %) of the total uncharacterized genes in CGD (Figure 3). As

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

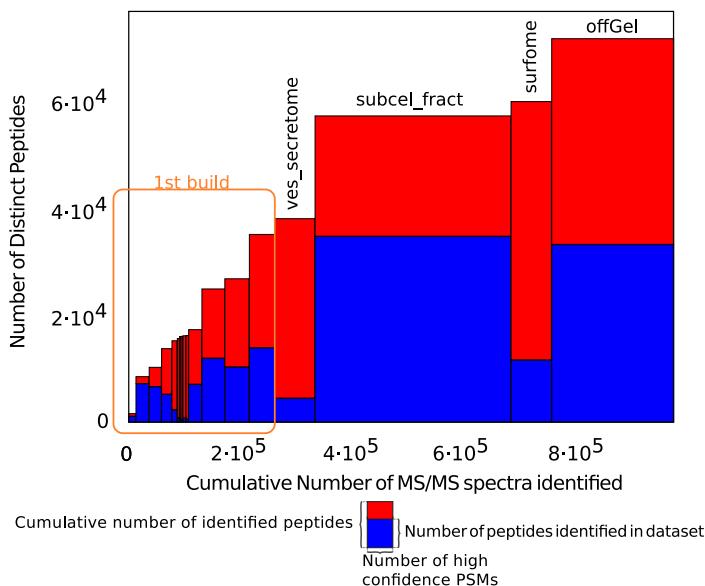


Figure 2. Contribution of the different constituent datasets of the *C. albicans* PeptideAtlas. The two implemented *ad hoc* strategies (subcel_fract and OFFGEL, as named in the web interface) have both the largest numbers of distinct peptides (height of blue bars) and contribute most to the increasing of the cumulative number of distinct peptides identified (total height of bars), and also represent the largest contributions in terms of spectra identified (width of bars). The experiments that constituted the previous PeptideAtlas are annotated for comparison.

for the verified set of genes, those that do have experimental evidence for a gene product, 76 % are covered in the list of canonical proteins in this build. (See Supplementary Tables CGD_uncharacterized_vs_PA_canonical.xls and CGD_verified_vs_PA_canonical.xls)

Gene Ontology enrichment analysis of the covered and undetected proteome subsets

The set of 4174 identified canonical proteins was subjected to a GO term enrichment analysis using GOSlimMapper in CGD. This analysis revealed no specific bias towards any particular biological process in the covered part of the proteome showing a very similar (no statistically significant difference) histogram of frequencies of GO Slim biological process terms as that for the entire genome at CGD.

The undetected set of 2103 proteins, obtained by subtracting the 4115 canonical proteins

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

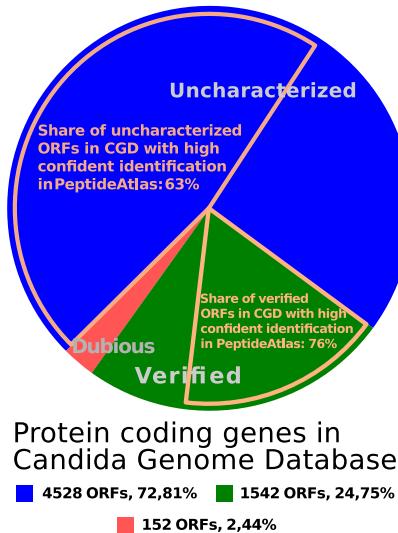


Figure 3. Share of the uncharacterized genes (genes for which there is no empirical evidence of a protein product) and Verified genes (those having a protein product with a given GO annotation) in CGD that are covered by canonical (with high confidence identification) proteins in the *C. albicans* PeptideAtlas.

in PeptideAtlas from the 6218 predicted proteins in CGD, was similarly enriched in GO Slim biological processes revealing, as expected, very heterogeneous annotations with a majority of the genes that cannot be grouped under a more precise category (in the Slim pruned GO tree) than "biological process" which means these are likely uncharacterized genes. This undetected set is where the focus should be laid on to further extend the proteome coverage in future builds of the *C. albicans* PeptideAtlas by designing specific strategies to detect, at least, those proteins that do have some specific biological process or cellular component annotations.

Both GO analyses for the covered and undetected subsets are available in supplementary material (GO_SLIM_PA_201503.xls and GO_SLIM_undetectedCGD_201503.xls)

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

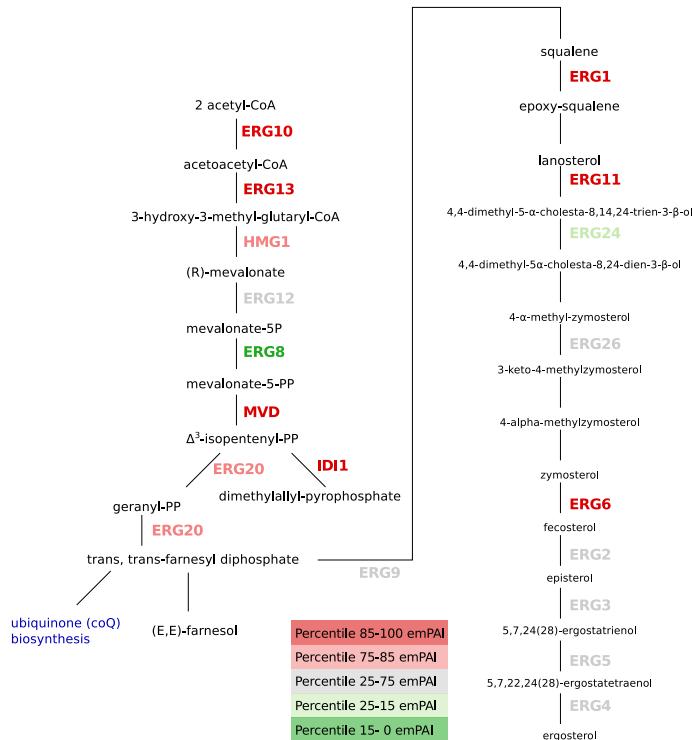


Figure 4. The *C. albicans* PeptideAtlas contains information on the detection and estimated abundances (emPAI) of proteins representing 18 out of 22 steps in the biosynthesis of ergosterol, an essential pathway comprising targets of many antifungal drugs.

Assessment of protein abundance and functional analysis

Once the abundance clusters were established based on the emPAI method for the set of 3,000 proteins identified in the subcellular fractionation approach (supplementary file *emPAI_results.xls*), a functional analysis was carried out on them. First, they were mapped onto the ergosterol biosynthesis pathway (Figure 4), which is of great interest since it is specific to fungi (the functional equivalent in mammalian cell membranes is cholesterol) and is therefore the target of many antifungal drugs that exploit selective toxicity. In addition, farnesol, a by-product in this pathway, has been shown to have a role in quorum sensing (Albuquerque and

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

Casadevall, 2012) and apoptosis induction (Léger et al., 2015). As depicted in Figure 4, most of the proteins, representing 18 out of 22 steps in the pathway, were detected, with a majority of them belonging in the high abundance groups.

Then, a GO enrichment analysis was also applied to the abundance sets, in this case combining the two high abundance protein groups on one hand, and the two low abundance groups on the other. The top 3 enriched and under-represented GO-Slim annotations of each the Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) ontologies were selected and shown in figure 5. Interestingly, low abundance proteins appear to be enriched in the BP terms *pseudohyphal growth* and *cell budding* and consistently in *hyphal tip* and *site of polarized growth* CC annotations. Signal transduction and kinase activities are also enriched in this set of low abundance proteins, in agreement with the described low quantities of the proteins that carry out these functions. The set of high abundance proteins, as expected, are enriched in some of the terms for which the low abundance proteins are under-represented, such as *cell wall* and *extracellular region*, and conversely are under-represented in some other processes and functions in which low abundance proteins seem to be involved, like *pseudohyphal growth* and *kinase activity*.

Allele specific proteins

Taking advantage of the database containing allele specific sequences, we have examined the lists of proteins that can be allocated to their specific originating allele. Of the 4174 identified canonical *C. albicans* protein sequences, there are 59 pairs of alleles with different sequences for which both protein products have been identified through their differentiating peptides.

There are 354 proteins from allele B that have been labeled canonical without a similar detection of their corresponding allele A counterparts. This means there is solid evidence for the presence of the protein originating from allele B, but does not necessarily imply that only the form from allele B was present. Proteins from allele A might have also been present in the samples but are not included in the minimal non-redundant list either because all their identified peptides are shared and can be explained by the canonical B which do have additional exclusive peptide evidence (the A protein forms are subsumed, in PeptideAtlas terminology); or because they have a too weak peptide evidence that could distinguish them from the canonical identification (termed possibly distinguished and conservatively excluded from the canonical list). Three more proteins (C5_01745W_B, C5_01765C_B and C5_01775C_B) from allele B were identified but are related to the mating type locus (MTL), and therefore exist only in haplotype B. And 11 more proteins, encoded by mitochondrial DNA, cannot be allocated to either allele. Subtracting the canonical proteins from allele B and the described particular cases from the 4056 protein count that excludes identifications from both alleles, there remain 3688 (4056

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

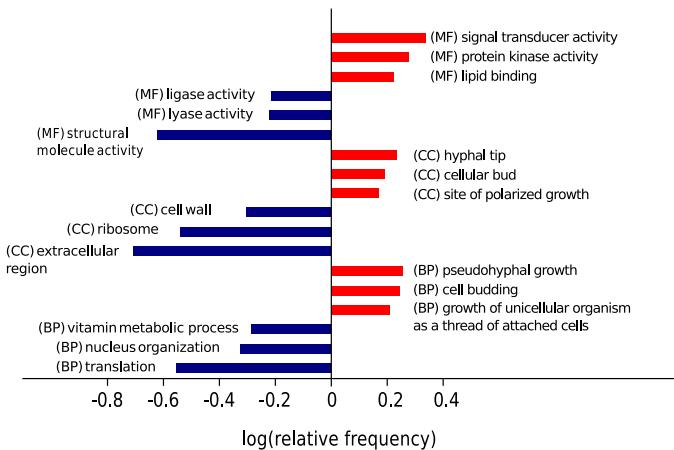
| Allele A | Allele B | Sequences A and B | mtDNA |
|---|--------------------------------------|-------------------|--------------|
| 59 canonical | 59 canonical | different | - |
| subsumed or possibly distinguished | 354 canonical | different | - |
| - | 3 canonical from MTL exist only in B | - | - |
| - | - | - | 11 canonical |
| 2070 canonical (chosen as representative) | identical | identical | - |
| 1205 canonical | indistinguishable | different | - |
| 413 canonical | subsumed or possibly distinguished | different | - |

Table 2. Distribution of the canonical proteins in PeptideAtlas by genomic origin. In the PeptideAtlas terminology, subsumed refers to proteins whose peptides are also present in another canonical protein which has additional independent peptide evidence; possibly distinguished means a weak peptide evidence that could possibly distinguish the protein from the canonical identification (these are conservatively excluded from the canonical list); and indistinguishable refers to proteins having different sequence but with peptide evidence only in the common parts.

- 354 - 3 - 11) identifications that could be originated from allele A. However, within these, 2070 have identical allele A and B sequences. In those cases, either allele is equally likely the origin of the identified protein and any one allele can be chosen as representative. Notably, this does not imply that the remaining 1618 (3688 - 2070) proteins necessarily correspond exclusively to allele A. Out of these, 1205 allele B forms are indistinguishable to that from A, which means their identified peptides are the same and mapped to common parts of the protein sequence. And lastly, 413 do have independent peptide evidence of being originated from allele A, but yet again this is not exclusive, the form from allele B might be subsumed or possibly distinguished. An overview of how the canonical proteins in the *C. albicans* PeptideAtlas are distributed by genomic origin and the presence level is summarized in Table 2.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES
DEEP PROTEOME COVERAGE

a) Very low and low abundance proteins
Relative frequencies of GO Slim annotations



b) Very high and high abundance proteins
Relative frequencies of GO Slim annotations

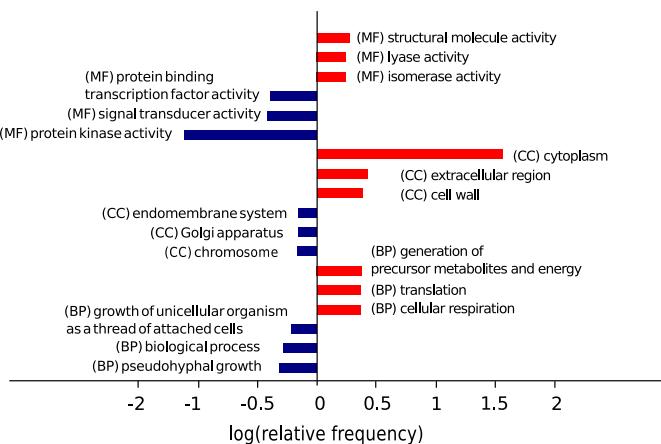


Figure 5. Top three enriched and under-represented annotations for the Molecular Function, Cellular Component and Biological Process ontologies for the sets of very low and low abundance proteins (a) and high and very high abundance proteins (b) in the subcellular fractionation approach.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

Conclusions

We have described the new *C. albicans* PeptideAtlas 2015-02 which represents a great increase in the number of characterized peptides and proteins with respect to the previous build. A total of 71,310 peptides and 4174 protein sequences make it the most comprehensive proteomics resource available up to date with a coverage of 66 % of the total predicted proteome. In addition, highly confident protein identifications have been reported for 63 % of the genes termed uncharacterized (without a known protein product) in CGD.

Furthermore, for the first time in a large-scale *C. albicans* proteomics project, an allele-specific protein sequence database has been searched and integrated into the resource enabling the ability to trace the identified proteins back to their originating allele. This, for example, enables the development of targeted assays to distinguish protein isoforms via the PeptideAtlas web interface (Farrah et al., 2011) to select candidate proteotypic peptides for the basis of the best peptides to use.

While this effort provides an unbiased representative picture of the whole *C. albicans* proteome, there is still room for further improvement. Future PeptideAtlas builds may include datasets specifically generated to detect the elusive proteins that may be expressed only under very particular circumstances, difficult to extract proteins, or may be translated in very low quantities. Finally, improvements in the software pipeline used for post-acquisition analysis or in the protein sequence database will also motivate the construction of new *C. albicans* PeptideAtlas builds in the future.

References

- Albuquerque, P. and Casadevall, A. (2012), Quorum sensing in fungi - a review., Medical mycology, 50(4), 337-45.
- Cabezón, V., Llama-Palacios, A., Nombela, C., Monteoliva, L., and Gil, C. (2009), Analysis of *Candida albicans* plasma membrane proteome., Proteomics, 9(20), 4770-86.
- Cheng, S.-C., Joosten, L. a. B., Kullberg, B.-J., and Netea, M. G. (2012), Interplay between *Candida albicans* and the mammalian innate host defense., Infection and immunity, 80(4), 1304-13.
- Costanzo, M. C., Arnaud, M. B., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., and Sherlock, G. (2006), The *Candida* Genome Database: facilitating research on *Candida albicans* molecular biology., FEMS yeast research, 6(5), 671-84.
- Craig, R. and Beavis, R. C. (2004), TANDEM: matching proteins with tandem mass spectra., Bioinformatics, 20(9), 1466-7.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

- Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006), The PeptideAtlas project., Nucleic acids research, 34(Database issue), D655-8.
- Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., and Moritz, R. L. (2015), Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics., Proteomics. Clinical applications.
- Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013), Comet: an open-source MS/MS sequence database search tool., Proteomics, 13(1), 22-4.
- Fernández-Arenas, E., Cabezón, V., Bermejo, C., Arroyo, J., Nombela, C., Diez-Orejas, R., and Gil, C. (2007), Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction., Molecular & cellular proteomics: MCP, 6(3), 460-478.
- Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004), Open mass spectrometry search algorithm., Journal of proteome research, 3(5), 958-64.
- Gil-Bona, A., Llama-Palacios, A., Parra, C. M., Vivanco, F., Nombela, C., Monteoliva, L., and Gil, C. (2014), Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in *Candida albicans*., Journal of proteome research, 14(1), 142-53.
- Gil-Bona, A., Parra-Giraldo, C. M., Hernández, M. L., Reales-Calderon, J. A., Solis, N. V., Fi-ller, S. G., Monteoliva, L., and Gil, C. (2015), *Candida albicans* cell shaving uncovers new proteins involved in cell wall integrity, yeast to hypha transition, stress response and host-pathogen interaction., Journal of proteomics.
- Gow, N. and van de Veerdonk, F. (2011), *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization, Nature Reviews, 10(2), 112-122.
- Gunaratne, J., Schmidt, A., Quandt, A., Neo, S. P., Sarac, O. S., Gracia, T., Loguercio, S., Ahrne, E., Li Hai Xia, R., Tan, K. H., Loessner, C., Bahler, J., Beyer, A., Blackstock, W., and Aebersold, R. (2013), Extensive Mass Spectrometry-Based Analysis of the Fission Yeast Proteome: The *S. pombe* PeptideAtlas, Molecular & Cellular Proteomics, pp. 1741-1751.
- Hao, B., Cheng, S., Clancy, C. J., and Nguyen, M. H. (2013), Caspofungin kills *Candida albicans* by causing both cellular apoptosis and necrosis., Antimicrobial agents and chemotherapy, 57(1), 326-32.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

Harford, J. B. and Bonifacino, J. S. (2001), Subcellular Fractionation and Isolation of Organelles, in Current Protocols in Cell Biology. John Wiley & Sons, Inc.

Horth, P., Miller, C. A., Preckel, T., and Wenz, C. (2006), Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis., Molecular & cellular proteomics : MCP, 5(10), 1968-74.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappaport, J., and Mann, M. (2005), Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein., Molecular & cellular proteomics : MCP, 4(9), 1265-72.

Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002), Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search., Analytical chemistry, 74(20), 5383-92.

Keller, A., Eng, J., Zhang, N., Li, X.-J., and Aebersold, R. (2005), A uniform proteomics MS/MS analysis platform utilizing open XML file formats, Molecular systems biology, 1(August 2005), 2005.0017.

King, N. L., Deutsch, E. W., Ranish, J. A., Nesvizhskii, A. I., Eddes, J. S., Mallick, P., Eng, J., Desiere, F., Flory, M., Martin, D. B., Kim, B., Lee, H., Raught, B., and Aebersold, R. (2006), Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas., Genome biology, 7(11), R106

Léger, T., Garcia, C., Ounissi, M., Lelandais, G., and Camadro, J.-M. (2015), The metacaspase (Mca1p) has a dual role in farnesol-induced apoptosis in *Candida albicans*., Molecular & cellular proteomics : MCP, 14(1), 93-108.

MacLean, B., Eng, J., Beavis, R., and McIntosh, M. (2006), General framework for developing and evaluating database scoring algorithms using the TANDEM search engine, Bioinformatics, 22(22), 2830-2832.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Rompp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011), mzML-a community standard for mass spectrometry data., Molecular & cellular proteomics : MCP, 10(1), R110.000133.

Monteoliva, L., Martinez-Lopez, R., Pitarch, A., Hernaez, M. L., Serna, A., Nombela, C., Albar, J. P., and Gil, C. (2011), Quantitative proteome and acidic subproteome profiling

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES DEEP PROTEOME COVERAGE

of *Candida albicans* yeast-to-hypha transition, *Journal of Proteome Research*, 10(2), 502-517.

Muzzey, D., Schwartz, K., Weissman, J. S., and Sherlock, G. (2013), Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure., *Genome biology*, 14(9), R97.

Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003), A statistical model for identifying proteins by tandem mass spectrometry., *Analytical chemistry*, 75(17), 4646-58.

Pitarch, A., Nombela, C., and Gil, C. (2006a), *Candida albicans* biology and pathogenicity: insights from proteomics., *Methods of biochemical analysis*, 49, 285-330.

Pitarch, A., Nombela, C., and Gil, C. (2009), Proteomic profiling of serologic response to *Candida albicans* during host-commensal and host-pathogen interactions., in *Methods in molecular biology* (Clifton, N.J.), vol. 470, pp. 369-411.

Pitarch, A., Nombela, C., and Gil, C. (2011), Prediction of the clinical outcome in invasive candidiasis patients based on molecular fingerprints of five anti-*Candida* antibodies in serum., *Molecular & Cellular Proteomics*, 10(1), M110.004010.

Ramsdale, M. (2008), Programmed cell death in pathogenic fungi, *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1783(7), 1369-1380.

Ros, A., Faupel, M., Mees, H., van Oostrum, J., Ferrigno, R., Reymond, F., Michel, P., Rossier, J. S., and Girault, H. H. (2002), Protein purification by Off-Gel electrophoresis., *Proteomics*, 2(2), 151-6.

Rupp, S. (2004), Proteomics on its way to study host- pathogen interaction in *Candida albicans*, *Current opinion in microbiology*, 7(4), 330-335.

Shinoda, K., Tomita, M., and Ishihama, Y. (2010), emPAI Calc - for the estimation of protein abundance from large-scale identification data by liquid chromatography - tandem mass spectrometry., *Bioinformatics (Oxford, England)*, 26(4), 576-7.

Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, a. I. (2011), iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates, *Molecular & Cellular Proteomics*, 10(12), M111.007690-M111.007690.

A COMPREHENSIVE *Candida albicans* PEPTIDEATLAS BUILD ENABLES
DEEP PROTEOME COVERAGE

Vialás, V., Perumal, P., Gutierrez, D., Ximénez-Embún, P., Nombela, C., Gil, C., and Chaffin, W. L. (2012), Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells., *Proteomics*, 12(14), 2331-2339.

Vialas, V., Sun, Z., Loureiro Y Penha, C. V., Carrascal, M., Abiñan, J., Monteoliva, L., Deutsch, E. W., Aebersold, R., Moritz, R. L., and Gil, C. (2013), A *Candida albicans* PeptideAtlas., *Journal of proteomics*, 97, 62-8.

DISCUSIÓN

Discusión

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

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

DISCUSIÓN

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

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

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

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

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

Con la adopción del formato mzIdentML (version 1.1.0) como fuente de resultados, Proteopathogen ha adquirido la capacidad de crecer de manera robusta y fiable. Por una parte, usar un único tipo de formato, estándar, como fuente de datos ha permitido desarrollar un software estable que extrae la información independientemente de cómo se hayan obtenido los resultados. Pero además, el uso de este estándar permite realizar validaciones, un control de calidad, tanto de la estructura, sintaxis y orden de los elementos XML de los archivos (validación semántica), como del contenido mínimo (validación MIAPE) usando para ello algunas herramientas como la creada por HUPO-PSI, mzidValidator (Ghali et al., 2013).

DISCUSIÓN

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

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

Otro tipo de posible mejora en la aplicación podría consistir en implementar un nuevo modo de leer los archivos mzIdentML. Esta posibilidad vendría motivada por la creciente capacidad de adquisición de datos de los espectrómetros de masas y las mejoras en el análisis bioinformático subsiguiente, que se traduce en archivos de resultados que llegan a tener un gran tamaño para archivos de texto (hasta el orden de gigabytes). El modo en que Proteopathogen lee los archivos consiste en una representación en memoria de toda la jerarquía xml (*parser DOM*). Esto, que es muy efectivo para localizar los distintos elementos referenciados en el contenido y guardarlos en cada tabla correspondiente de la base de datos, puede convertirse en una tarea difícil (requerir computadores con gran memoria de trabajo, RAM) o incluso imposible en algunos casos. Por este motivo puede ser interesante explorar otro tipo de implementaciones de lectura de archivos XML (*parser SAX*) que permitan leer la información contenida en archivos de gran tamaño.

DISCUSIÓN

Desarrollo de un PeptideAtlas de *Candida albicans*

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

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

Pero PeptideAtlas permite y promueve que los espectros sean reprocesados cuando se obtienen nuevos resultados de LC-MS/MS o cuando se desarrollan nuevas mejoras en los motores de búsqueda o en el *software* de análisis. Así, trás el PeptideAtlas original se obtuvieron nuevos resultados experimentales. Algunos de ellos fueron específicamente destinados a la ampliación de la cobertura del proteoma mediante fraccionamientos extensivos a varios niveles: subcelular (centrifugaciones), proteína (SDS-PAGE) y péptido (OFF-GEL); mientras que otros eran procedentes de trabajos destinados al estudio de proteínas de la pared celular (Gil-Bona et al., 2015) o secretadas mediante vesículas (Gil-Bona et al., 2014). Además en ese tiempo apareció un nuevo ensamblaje de la secuencia de *C. albicans* que por primera vez incluía secuencias específicas de alelo. (Muzzey et al., 2013). En este contexto, con los nuevos sets de datos y la nueva información de secuencia disponible, el PeptideAtlas original ha sido reanalizado, en conjunto con los sets de datos que formaban la primera versión, y empleando tres motores de búsqueda (SEQUEST, OMSSA y Comet) para finalmente obtener una cobertura de dos terceras partes del proteoma predicho de *C. albicans*. Más de 71.000 péptidos asignados a 4174 secuencias de proteínas (para un FDR de 0.10 % a nivel de PSM) suponen exactamente 66.17 % del proteoma, y con respecto a la versión inicial, un incremento de 3 veces el número de péptidos y 1.6 veces el de proteínas.

Con este notable incremento, este PeptideAtlas continúa siendo el recurso público de datos proteómicos de *C. albicans* más exhaustivo. Pero además de describir una lista de proteínas que representa un 66 % del proteoma predicho en CGD (*Candida Genome Database*), el PeptideAtlas de *C. albicans* supone una novedad y un gran valor añadido al proporcionar evi-

DISCUSIÓN

dencia empírica sólida de la existencia de proteínas para dos terceras partes de los genes que en CGD se denominan *uncharacterized* por carecer de un producto proteico caracterizado.

Además este PeptideAtlas es el gran primer repositorio *online* de Proteómica de *C. albicans* que permite mantener la trazabilidad desde la identificación de péptidos y proteínas hasta el alelo original.

Una aplicación de gran utilidad que permite la interfaz web consiste en facilitar la selección de péptidos proteotípicos candidatos para ensayos de Proteómica dirigida (SRM/MRM). Para ello, en la web se sugieren péptidos con una única localización genómica, lo que permite asegurar que sean exclusivos de una sola proteína y no compartidos. Además, los péptidos se ordenan por un índice de observabilidad que indica la probabilidad de que una proteína sea detectada por medio de esos péptidos y no otros (Deutsch et al., 2008b).

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

CONCLUSIONES

Conclusiones

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

Bibliografía

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

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

Abdi, H. H. (2007), 'The Bonferroni and Sidak Corrections for Multiple Comparisons', *Encyclopedia of Measurement and Statistics.*, **1**, 1–9.

Albuquerque, P. and Casadevall, A. (2012), 'Quorum sensing in fungi—a review.', *Medical mycology*, **50**(4), 337–45.

Arnaud, M. B., Costanzo, M. C., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., and Sherlock, G. (2005), 'The Candida Genome Database (CGD), a community resource for Candida albicans gene and protein information.', *Nucleic acids research*, **33**(Database issue), D358–63.

Barsnes, H., Vaudel, M., Colaert, N., Helsens, K., Sickmann, A., Berven, F. S., and Martens, L. (2011), 'compomics-utilities: an open-source Java library for computational proteomics.', *BMC bioinformatics*, **12**, 70.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000), 'The Protein Data Bank.', *Nucleic acids research*, **28**(1), 235–42.

Berman, J. and Sudbery, P. E. (2002), 'Candida Albicans: a molecular revolution built on lessons from budding yeast.', *Nature reviews. Genetics*, **3**(12), 918–30.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Bislev, S., Deutsch, E., and Sun, Z. (2012), 'A Bovine PeptideAtlas of milk and mammary gland proteomes', *Molecular & Cellular Proteomics*, **12**(18), 2895–2899.
- Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., and Blake, J. A. (2008), 'The Mouse Genome Database (MGD): mouse biology and model systems.', *Nucleic acids research*, **36**(Database issue), D724–8.
- Cabezón, V., Llama-Palacios, A., Nombela, C., Monteoliva, L., and Gil, C. (2009), 'Analysis of *Candida albicans* plasma membrane proteome.', *Proteomics*, **9**(20), 4770–86.
- Calderone, R. (2012), *Candida and candidiasis*. ASM Press, Washington DC.
- Castillo, L., Calvo, E., Martínez, A. I., Ruiz-Herrera, J., Valentín, E., Lopez, J. A., and Sentandreu, R. (2008), 'A study of the *Candida albicans* cell wall proteome.', *Proteomics*, **8**(18), 3871–81.
- Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Freywen, B., Baker, T. A., Brusniak, M.-Y., Pausle, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012), 'A cross-platform toolkit for mass spectrometry and proteomics.', *Nature biotechnology*, **30**(10), 918–20.
- Chan, Q. W. T., Parker, R., Sun, Z., Deutsch, E. W., and Foster, L. J. (2011), 'A honey bee (*Apis mellifera* L.) PeptideAtlas crossing castes and tissues.', *BMC genomics*, **12**(1), 290.
- Cheng, S.-C., Joosten, L. a. B., Kullberg, B.-J., and Netea, M. G. (2012), 'Interplay between *Candida albicans* and the mammalian innate host defense.', *Infection and immunity*, **80**(4), 1304–13.
- Choi, H. and Nesvizhskii, A. I. (2008), 'Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics.', *Journal of proteome research*, **7**(1), 254–65.
- Costanzo, M. C., Arnaud, M. B., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., and Sherlock, G. (2006), 'The *Candida* Genome Database: facilitating research on *Candida albicans* molecular biology.', *FEMS yeast research*, **6**(5), 671–84.
- Craig, R. and Beavis, R. C. (2004), 'TANDEM: matching proteins with tandem mass spectra.', *Bioinformatics*, **20**(9), 1466–7.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Csank, C., Costanzo, M. C., Hirschman, J., Hodges, P., Kranz, J. E., Mangan, M., O'Neill, K., Robertson, L. S., Skrzypek, M. S., Brooks, J., and Garrels, J. I. (2002), 'Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD).', *Methods in enzymology*, **350**, 347–73.
- de Groot, P. W. J., Bader, O., de Boer, A. D., Weig, M., and Chauhan, N. (2013), 'Adhesins in human fungal pathogens: glue with plenty of stick.', *Eukaryotic cell*, **12**(4), 470–81.
- Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006), 'The PeptideAtlas project.', *Nucleic acids research*, **34**(Database issue), D655–8.
- Deutsch, E. (2012), 'File formats commonly used in mass spectrometry proteomics', *Molecular & Cellular Proteomics*, **11**(12), 1612–1621.
- Deutsch, E., Lam, H., and Aebersold, R. (2008a), 'Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics', *Physiological genomics*, **33**(1), 18–25.
- Deutsch, E. E. W., Lam, H., and Aebersold, R. (2008b), 'PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows', *EMBO reports*, **9**(5), 429–434.
- Deutsch, E. E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., and Aebersold, R. (2010), 'A guided tour of the TransProteomic Pipeline', *Proteomics*, **10**(6), 1150–1159.
- Deutsch, E. W. (2010), 'Mass spectrometer output file format mzML.', *Methods in molecular biology (Clifton, N.J.)*, **604**, 319–31.
- Deutsch, E. W., Chambers, M., Neumann, S., Levander, F., Binz, P.-A., Shofstahl, J., Campbell, D. S., Mendoza, L., Ovelleiro, D., Helsens, K., Martens, L., Aebersold, R., Moritz, R. L., and Brusniak, M.-Y. (2012), 'TraML—a standard format for exchange of selected reaction monitoring transition lists.', *Molecular & cellular proteomics : MCP*, **11**(4), R111.015040.
- Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., and Moritz, R. L. (2015), 'Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics.', *Proteomics. Clinical applications*.
- Ding, C., Chan, D. W., Liu, W., Liu, M., Li, D., Song, L., Li, C., Jin, J., Malovannaya, A., Jung, S. Y., Zhen, B., Wang, Y., and Qin, J. (2013), 'Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements', *Proceedings of the National Academy of Sciences of the United States of America*, **110**(17), 6771–6.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Ding, Y., Choi, H., and Nesvizhskii, A. I. (2008), 'Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics.', *Journal of proteome research*, **7**(11), 4878–89.
- Elias, J. E. and Gygi, S. P. (2007), 'Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.', *Nature methods*, **4**(3), 207–14.
- Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013), 'Comet: an open-source MS/MS sequence database search tool.', *Proteomics*, **13**(1), 22–4.
- Eng, J. K., McCormack, A. L., and Yates, J. R. (1994), 'An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.', *Journal of the American Society for Mass Spectrometry*, **5**(11), 976–89.
- Fanning, S. and Mitchell, A. P. (2012), 'Fungal biofilms.', *PLoS pathogens*, **8**(4), e1002585.
- Farrah, T., Deutsch, E. W., Omenn, G. S., Campbell, D. S., Sun, Z., Bletz, J. a., Mallick, P., Katz, J. E., Malmström, J., Ossola, R., Watts, J. D., Lin, B., Zhang, H., Moritz, R. L., and Aebersold, R. (2011), 'A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas.', *Molecular & Cellular Proteomics*, **10**(9), M110.006353.
- Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989), 'Electrospray ionization for mass spectrometry of large biomolecules.', *Science (New York, N.Y.)*, **246**(4926), 64–71.
- Fenyö, D. and Beavis, R. (2003), 'A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes', *Analytical Chemistry*, **75**(4), 768–774.
- Fernández-Arenas, E., Cabezón, V., Bermejo, C., Arroyo, J., Nombela, C., Diez-Orejas, R., and Gil, C. (2007), 'Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction.', *Molecular & cellular proteomics : MCP*, **6**(3), 460–478.
- Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004), 'Open mass spectrometry search algorithm.', *Journal of proteome research*, **3**(5), 958–64.
- Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003), 'Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.', *Proceedings of the National Academy of Sciences of the United States of America*, **100**(12), 6940–5.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Ghali, F., Krishna, R., Lukasse, P., Martínez-Bartolomé, S., Reisinger, F., Hermjakob, H., Vizcaíno, J. A., and Jones, A. R. (2013), 'Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML.', *Molecular & cellular proteomics : MCP*, **12**(11), 3026–35.
- Gil-Bona, A., Llama-Palacios, A., Parra, C. M., Vivanco, F., Nombela, C., Monteoliva, L., and Gil, C. (2014), 'Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in *Candida albicans*.', *Journal of proteome research*, **14**(1), 142–53.
- Gil-Bona, A., Parra-Giraldo, C. M., Hernández, M. L., Reales-Calderon, J. A., Solis, N. V., Filler, S. G., Monteoliva, L., and Gil, C. (2015), 'Candida albicans cell shaving uncovers new proteins involved in cell wall integrity, yeast to hypha transition, stress response and host-pathogen interaction.', *Journal of proteomics*.
- Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012), 'Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.', *Molecular & Cellular Proteomics*, **11**(6), O111.016717.
- Gow, N. and van de Veerdonk, F. (2011), 'Candida albicans morphogenesis and host defence: discriminating invasion from colonization', *Nature Reviews*, **10**(2), 112–122.
- Griss, J., Reisinger, F., Hermjakob, H., and Vizcaíno, J. A. (2012), 'jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats.', *Proteomics*, **12**(6), 795–8.
- Gunaratne, J., Schmidt, A., Quandt, A., Neo, S. P., Sarac, O. S., Gracia, T., Loguercio, S., Ahrne, E., Li Hai Xia, R., Tan, K. H., Loessner, C., Bahler, J., Beyer, A., Blackstock, W., and Aebersold, R. (2013), 'Extensive Mass Spectrometry-Based Analysis of the Fission Yeast Proteome: The *S. pombe* PeptideAtlas', *Molecular & Cellular Proteomics*, pp. 1741–1751.
- Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000), 'Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.', *Proceedings of the National Academy of Sciences of the United States of America*, **97**(17), 9390–5.
- Hao, B., Cheng, S., Clancy, C. J., and Nguyen, M. H. (2013), 'Caspofungin kills *Candida albicans* by causing both cellular apoptosis and necrosis.', *Antimicrobial agents and chemotherapy*, **57**(1), 326–32.
- Harford, J. B. and Bonifacino, J. S. (2001), 'Subcellular Fractionation and Isolation of Organelles', in *Current Protocols in Cell Biology*. John Wiley & Sons, Inc.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Hatt, P. D., Quadroni, M., Staudenmann, W., and James, P. (1997), 'Concentration of, and SDS Removal from Proteins Isolated from Multiple two-Dimensional Electrophoresis Gels', *European Journal of Biochemistry*, **246**(2), 336–343.
- Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993), 'Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases.', *Proceedings of the National Academy of Sciences of the United States of America*, **90**(11), 5011–5.
- Hoehamer, C. F., Cummings, E. D., Hilliard, G. M., and Rogers, P. D. (2010), 'Changes in the proteome of *Candida albicans* in response to azole, polyene, and echinocandin antifungal agents.', *Antimicrobial agents and chemotherapy*, **54**(5), 1655–64.
- Hoogland, C., Mostaguir, K., Appel, R. D., and Lisacek, F. (2008), 'The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server.', *Journal of proteomics*, **71**(2), 245–8.
- Hörth, P., Miller, C. A., Preckel, T., and Wenz, C. (2006), 'Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis.', *Molecular & cellular proteomics : MCP*, **5**(10), 1968–74.
- Insenser, M., Nombela, C., Molero, G., and Gil, C. (2006), 'Proteomic analysis of detergent-resistant membranes from *Candida albicans*', *Proteomics*, **6 Suppl 1**, S74–81.
- Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005), 'Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.', *Molecular & cellular proteomics : MCP*, **4**(9), 1265–72.
- Jacobsen, I. D., Wilson, D., Wächtler, B., Brunke, S., Naglik, J. R., and Hube, B. (2012), 'Candida albicans dimorphism as a therapeutic target.', *Expert review of anti-infective therapy*, **10**(1), 85–93.
- Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S. J., Selley, J. N., Searle, B. C., Shofstahl, J., Seymour, S. L., Julian, R., Binz, P.-A., Deutsch, E. W., Hermjakob, H., Reisinger, F., Griss, J., Vizcaíno, J. A., Chambers, M., Pizarro, A., and Creasy, D. (2012), 'The mzIdentML data standard for mass spectrometry-based proteomics results.', *Molecular & cellular proteomics : MCP*, **11**(7), M111.014381.
- Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008), 'Assigning significance to peptides identified by tandem mass spectrometry using decoy databases.', *Journal of proteome research*, **7**(1), 29–34.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2007), 'KEGG for linking genomes to life and the environment', *Nucleic Acids Research*, **36**(Database), D480–D484.
- Karas, M. and Hillenkamp, F. (1988), 'Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.', *Analytical chemistry*, **60**(20), 2299–301.
- Keller, A., Eng, J., Zhang, N., Li, X.-j., and Aebersold, R. (2005), 'A uniform proteomics MS/MS analysis platform utilizing open XML file formats', *Molecular systems biology*, **1**(August 2005), 2005.0017.
- Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002), 'Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.', *Analytical chemistry*, **74**(20), 5383–92.
- King, N. L., Deutsch, E. W., Ranish, J. A., Nesvizhskii, A. I., Eddes, J. S., Mallick, P., Eng, J., Desiere, F., Flory, M., Martin, D. B., Kim, B., Lee, H., Raught, B., and Aebersold, R. (2006), 'Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas.', *Genome biology*, **7**(11), R106.
- Klose, J. (1975), 'Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals.', *Hu-mangenetik*, **26**(3), 231–43.
- Klose, J. and Kobalz, U. (1995), 'Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome.', *Electrophoresis*, **16**(6), 1034–59.
- Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W., and Guild, B. (2004), 'Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and ¹³C-labeled peptide standards.', *Proteomics*, **4**(4), 1175–86.
- Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., and Aebersold, R. (2007), 'Development and validation of a spectral library searching method for peptide identification from MS/MS.', *Proteomics*, **7**(5), 655–67.
- Lange, V., Malmström, J. a., Didion, J., King, N. L., Johansson, B. P., Schäfer, J., Rameseder, J., Wong, C.-H., Deutsch, E. W., Brusniak, M.-Y., Bühlmann, P., Björck, L., Domon, B., and Aebersold, R. (2008), 'Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring.', *Molecular & Cellular Proteomics*, **7**(8), 1489–500.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Léger, T., Garcia, C., Ounissi, M., Lelandais, G., and Camadro, J.-M. (2015), 'The metacaspase (Mca1p) has a dual role in farnesol-induced apoptosis in *Candida albicans*', *Molecular & cellular proteomics : MCP*, **14**(1), 93–108.
- Lindner, S. E., Swearingen, K. E., Harupa, A., Vaughan, A. M., Sinnis, P., Moritz, R. L., and Kappe, S. H. I. (2013), 'Total and putative surface proteomics of malaria parasite salivary gland sporozoites.', *Molecular & Cellular Proteomics*, **12**(5), 1127–43.
- Loevenich, S. N., Brunner, E., King, N. L., Deutsch, E. W., Stein, S. E., Aebersold, R., and Hafen, E. (2009), 'The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation.', *BMC bioinformatics*, **10**, 59.
- Ma, K., Vitek, O., and Nesvizhskii, A. I. (2012), 'A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet.', *BMC bioinformatics*, **13 Suppl 1**(Suppl 16), S1.
- MacLean, B., Eng, J., Beavis, R., and McIntosh, M. (2006), 'General framework for developing and evaluating database scoring algorithms using the TANDEM search engine', *Bioinformatics*, **22**(22), 2830–2832.
- Madeo, F., Herker, E., and Wissing, S. (2004), 'Apoptosis in yeast', *Current opinion in microbiology*, **7**(6), 655–660.
- Makarov, A. (2000), 'Electrostatic axially harmonic orbital trapping, a high performance technique of mass analysis', *Analytical chemistry*, **72**(6), 1156–62.
- Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold, R. (2007), 'Computational prediction of proteotypic peptides for quantitative proteomics.', *Nature biotechnology*, **25**(1), 125–31.
- Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpf, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011), 'mzML—a community standard for mass spectrometry data.', *Molecular & cellular proteomics : MCP*, **10**(1), R110.000133.
- Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005), 'PRIDE: the proteomics identifications database.', *Proteomics*, **5**(13), 3537–45.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Martínez-Bartolomé, S., Binz, P.-A., and Albar, J. P. (2014), 'The Minimal Information about a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative.', *Methods in molecular biology (Clifton, N.J.)*, **1072**, 765–80.
- Martínez-Solano, L., Nombela, C., Molero, G., and Gil, C. (2006), 'Differential protein expression of murine macrophages upon interaction with *Candida albicans*', *Proteomics*, **6 Suppl 1**, S133–S144.
- Mayer, F. L., Wilson, D., and Hube, B. (2013), 'Candida albicans pathogenicity mechanisms.', *Virulence*, **4**(2), 119–28.
- Monteoliva, L. and Albar, J. P. (2004), 'Differential proteomics: an overview of gel and non-gel based approaches.', *Briefings in functional genomics & proteomics*, **3**(3), 220–39.
- Monteoliva, L., Martinez-Lopez, R., Pitarch, A., Hernaez, M. L., Serna, A., Nombela, C., Albar, J. P., and Gil, C. (2011), 'Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition', *Journal of Proteome Research*, **10**(2), 502–517.
- Moran, C., Grussemeyer, C. A., Spalding, J. R., Benjamin, D. K., and Reed, S. D. (2010), 'Comparison of costs, length of stay, and mortality associated with *Candida glabrata* and *Candida albicans* bloodstream infections.', *American journal of infection control*, **38**(1), 78–80.
- Muzzey, D., Schwartz, K., Weissman, J. S., and Sherlock, G. (2013), 'Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure.', *Genome biology*, **14**(9), R97.
- Navarro, P. and Vázquez, J. (2009), 'A refined method to calculate false discovery rates for peptide identification using decoy databases.', *Journal of proteome research*, **8**(4), 1792–6.
- Nesvizhskii, A. (2010), 'A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics', *Journal of proteomics*, **73**(11), 2092–2123.
- Nesvizhskii, A. I. (2007), 'Protein identification by tandem mass spectrometry and sequence database searching.', *Methods in molecular biology (Clifton, N.J.)*, **367**, 87–119.
- Nesvizhskii, A. I. and Aebersold, R. (2005), 'Interpretation of shotgun proteomic data: the protein inference problem.', *Molecular & Cellular Proteomics*, **4**(10), 1419–40.
- Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003), 'A statistical model for identifying proteins by tandem mass spectrometry.', *Analytical chemistry*, **75**(17), 4646–58.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- O'Farrell, P. H. (1975), 'High resolution two-dimensional electrophoresis of proteins.', *The Journal of biological chemistry*, **250**(10), 4007–21.
- Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007), 'Higher-energy C-trap dissociation for peptide modification analysis.', *Nature methods*, **4**(9), 709–12.
- Pappin, D. J., Hojrup, P., and Bleasby, A. J. (1993), 'Rapid identification of proteins by peptide-mass fingerprinting.', *Current biology : CB*, **3**(6), 327–32.
- Pedrioli, P. G. A., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004), 'A common open representation of mass spectrometry data and its application to proteomics research.', *Nature biotechnology*, **22**(11), 1459–66.
- Pfaller, M. A. and Diekema, D. J. (2010), 'Epidemiology of invasive mycoses in North America.', *Critical reviews in microbiology*, **36**(1), 1–53.
- Phan, Q. T., Myers, C. L., Fu, Y., Sheppard, D. C., Yeaman, M. R., Welch, W. H., Ibrahim, A. S., Edwards, J. E., and Filler, S. G. (2007), 'Als3 is a *Candida albicans* invasin that binds to cadherins and induces endocytosis by host cells.', *PLoS biology*, **5**(3), e64.
- Pitarch, A., Abian, J., Carrascal, M., Sánchez, M., Nombela, C., and Gil, C. (2004), 'Proteomics-based identification of novel *Candida albicans* antigens for diagnosis of systemic candidiasis in patients with underlying hematological malignancies.', *Proteomics*, **4**(10), 3084–106.
- Pitarch, A., Nombela, C., and Gil, C. (2006a), 'Candida albicans biology and pathogenicity: insights from proteomics.', *Methods of biochemical analysis*, **49**, 285–330.
- Pitarch, A., Nombela, C., and Gil, C. (2006b), 'Contributions of proteomics to diagnosis, treatment, and prevention of candidiasis.', *Methods of biochemical analysis*, **49**, 331–61.
- Pitarch, A., Nombela, C., and Gil, C. (2009), 'Proteomic profiling of serologic response to *Candida albicans* during host-commensal and host-pathogen interactions.', in *Methods in molecular biology (Clifton, N.J.)*, vol. 470, pp. 369–411.
- Pitarch, A., Nombela, C., and Gil, C. (2011), 'Prediction of the clinical outcome in invasive candidiasis patients based on molecular fingerprints of five anti-*Candida* antibodies in serum.', *Molecular & Cellular Proteomics*, **10**(1), M110.004010.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Plaine, A., Walker, L., Da Costa, G., Mora-Montes, H. M., McKinnon, A., Gow, N. A. R., Gailardin, C., Munro, C. A., and Richard, M. L. (2008), 'Functional analysis of *Candida albicans* GPI-anchored proteins: roles in cell wall integrity and caspofungin sensitivity.', *Fungal genetics and biology : FG & B*, **45**(10), 1404–14.
- Pleissner, K.-P., Eifert, T., Buettner, S., Schmidt, F., Boehme, M., Meyer, T. F., Kaufmann, S. H. E., and Jungblut, P. R. (2004), 'Web-accessible proteome databases for microbial research.', *Proteomics*, **4**(5), 1305–13.
- Prieto, G., Aloria, K., Osinalde, N., Fullaondo, A., Arizmendi, J. M., and Matthiesen, R. (2012), 'PAnalyzer: a software tool for protein inference in shotgun proteomics.', *BMC bioinformatics*, **13**, 288.
- Ramsdale, M. (2008), 'Programmed cell death in pathogenic fungi', *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, **1783**(7), 1369–1380.
- Reiter, L., Claassen, M., Schrimpf, S. S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009), 'Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry', *Molecular & Cellular Proteomics*, **8**(11), 2405–2417.
- Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R., Beck, M., Brusniak, M.-Y., Hengartner, M. O., and Aebersold, R. (2011), 'mProphet: automated data processing and statistical validation for large-scale SRM experiments.', *Nature methods*, **8**(5), 430–5.
- Roepstorff, P. and Fohlman, J. (1984), 'Proposal for a common nomenclature for sequence ions in mass spectra of peptides.', *Biomedical mass spectrometry*, **11**(11), 601.
- Rogowska-Wrzesinska, A., Le Bihan, M.-C., Thaysen-Andersen, M., and Roepstorff, P. (2013), '2D gels still have a niche in proteomics.', *Journal of proteomics*, **88**, 4–13.
- Ros, A., Faupel, M., Mees, H., van Oostrum, J., Ferrigno, R., Reymond, F., Michel, P., Rossier, J. S., and Girault, H. H. (2002), 'Protein purification by Off-Gel electrophoresis.', *Proteomics*, **2**(2), 151–6.
- Rossignol, T., Lechat, P., Cuomo, C., Zeng, Q., Moszer, I., and DÉnfert, C. (2008), 'CandidaDB: a multi-genome database for Candida species and related Saccharomycotina.', *Nucleic acids research*, **36**(Database issue), D557–61.
- Rupp, S. (2004), 'Proteomics on its way to study host- pathogen interaction in *Candida albicans*', *Current opinion in microbiology*, **7**(4), 330–335.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Saville, S. P., Lazzell, A. L., Monteagudo, C., and Lopez-Ribot, J. L. (2003), 'Engineered control of cell morphology in vivo reveals distinct roles for yeast and filamentous forms of *Candida albicans* during infection.', *Eukaryotic cell*, **2**(5), 1053–60.
- Schubert, O. T., Mouritsen, J., Ludwig, C., Röst, H. L., Rosenberger, G., Arthur, P. K., Claassen, M., Campbell, D. S., Sun, Z., Farrah, T., Gengenbacher, M., Maiolica, A., Kaufmann, S. H. E., Moritz, R. L., and Aebersold, R. (2013), 'The Mtb Proteome Library: A Resource of Assays to Quantify the Complete Proteome of *Mycobacterium tuberculosis*', *Cell host & microbe*, **13**(5), 602–12.
- Shinoda, K., Tomita, M., and Ishihama, Y. (2010), 'emPAI Calc—for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry.', *Bioinformatics (Oxford, England)*, **26**(4), 576–7.
- Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, a. I. (2011), 'iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates', *Molecular & Cellular Proteomics*, **10**(12), M111.007690–M111.007690.
- Simicevic, J., Schmid, A. W., Gilardoni, P. A., Zoller, B., Raghav, S. K., Krier, I., Gubelmann, C., Lisacek, F., Naef, F., Moniatte, M., and Deplancke, B. (2013), 'Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics.', *Nature methods, advance on*.
- Smith, B. E., Hill, J. A., Gjukich, M. A., and Andrews, P. C. (2011), 'Tranche distributed repository and ProteomeCommons.org.', *Methods in molecular biology (Clifton, N.J.)*, **696**, 123–45.
- Steen, H. and Mann, M. (2004), 'The ABC's (and XYZ's) of peptide sequencing', *Nature Reviews Molecular Cell Biology*, **5**(9), 699–711.
- Sundstrom, P., Balish, E., and Allen, C. M. (2002), 'Essential role of the *Candida albicans* transglutaminase substrate, hyphal wall protein 1, in lethal oroesophageal candidiasis in immunodeficient mice.', *The Journal of infectious diseases*, **185**(4), 521–30.
- Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004), 'Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.', *Proceedings of the National Academy of Sciences of the United States of America*, **101**(26), 9528–33.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012), 'GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics.', *Nucleic acids research*, **40**(Web Server issue), W478–83.
- Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988), 'Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry', *Rapid Communications in Mass Spectrometry*, **2**(8), 151–153.
- The Uniprot Consortium (2008), 'The universal protein resource (UniProt).', *Nucleic acids research*, **36**(Database issue), D190–5.
- Tong, K. B., Murtagh, K. N., Lau, C., and Seifeldin, R. (2008), 'The impact of esophageal candidiasis on hospital charges and costs across patient subgroups.', *Current medical research and opinion*, **24**(1), 167–74.
- Van, P. T., Schmid, A. K., King, N. L., Kaur, A., Pan, M., Whitehead, K., Koide, T., Facciotti, M. T., Goo, Y. A., Deutsch, E. W., Reiss, D. J., Mallick, P., and Baliga, N. S. (2008), 'Halobacterium salinarum NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage.', *Journal of proteome research*, **7**(9), 3755–64.
- Vialas, V. and Gil, C. (2015), 'Proteopathogen2, a database and web tool to store and display proteomics identification results in the mzIdentML standard', *EuPA Open Proteomics*.
- Vialás, V., Nogales-Cadenas, R., Nombela, C., Pascual-Montano, A., and Gil, C. (2009), 'Proteopathogen, a protein database for studying *Candida albicans*–host interaction.', *Proteomics*, **9**(20), 4664–8.
- Vialás, V., Perumal, P., Gutierrez, D., Ximénez-Eembún, P., Nombela, C., Gil, C., and Chaffin, W. L. (2012), 'Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells.', *Proteomics*, **12**(14), 2331–2339.
- Vialas, V., Sun, Z., Loureiro Y Penha, C. V., Carrascal, M., Abián, J., Monteoliva, L., Deutsch, E. W., Aebersold, R., Moritz, R. L., and Gil, C. (2013), 'A *Candida albicans* PeptideAtlas.', *Journal of proteomics*, **97**, 62–8.
- Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. a., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Pérez-Riverol, Y., Reisinger, F., Ríos, D., Wang, R., and Hermjakob, H. (2013), 'The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013.', *Nucleic acids research*, **41**(Database issue), D1063–9.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., and Humphery-Smith, I. (1995), 'Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*.', *Electrophoresis*, **16**(7), 1090–4.
- Wisplinghoff, H., Bischoff, T., Tallent, S. M., Seifert, H., Wenzel, R. P., and Edmond, M. B. (2004), 'Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study.', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, **39**(3), 309–17.
- Wolters, D. A., Washburn, M. P., and Yates, J. R. (2001), 'An automated multidimensional protein identification technology for shotgun proteomics.', *Analytical chemistry*, **73**(23), 5683–90.
- Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998), 'Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process', *Journal of the American Chemical Society*, **120**(13), 3265–3266.

Lista de acrónimos

| | |
|--------------|---|
| 1D-PAGE | <i>Monodimensional PoliAcrylamide Gel Electrophoresis</i> |
| 2D-PAGE | <i>Bidimensional PoliAcrylamide Gel Electrophoresis</i> |
| CGD | <i>Candida Genome Database</i> |
| CID | <i>Collision Induced Dissociation</i> |
| DDA | <i>Data Dependent Acquisition</i> |
| ECD | <i>Electron Capture Dissociation</i> |
| EM | <i>Expectation-Maximization</i> |
| ESI | <i>Electro-Spray Ionization</i> |
| ETD | <i>Electron Transfer Dissociation</i> |
| FAB | <i>Fast Atom Bombardment</i> |
| FD | <i>Field Desorption</i> |
| FDR | <i>False Discovery Rate</i> |
| FNR | <i>False Negative Rate</i> |
| FPR | <i>False Positive Rate</i> |
| FTICR | <i>Fourier Transform Ion Cyclotron Resonance</i> |
| FWHM | <i>Full Width at Half Mass</i> |
| HCD | <i>Higher Energy Collision Dissociation</i> |

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- HPLC *High Performance Liquid Chromatography*
- HUPO-PSI.... *Human Proteome Organization - Proteomics Standards Initiative*
- LIT..... *Linear Ion Trap*
- LTQ..... *Linear Trap Quadrupole*
- MALDI..... *Matrix Assisted Laser Desorption Ionization*
- MS/MS *Tandem Mass Spectrometry*
- MUDPIT..... *Multidimensional Protein Identification Technology*
- NIST..... *National Institute for Standards and Technology*
- NMC..... *Number of Missed Cleavages*
- NRS..... *Number of Replicate Spectra*
- NSE..... *Number of Sibling Experiments*
- NSI..... *Number of Sibling Ions*
- NSM..... *Number of Sibling Modifications*
- NSP..... *Number of Sibling Peptides*
- NSS..... *Number of Sibling Searches*
- NTT..... *Number of Tryptic Termini*
- OMSSA..... *Open Mass Spectrometry Search Algorithm*
- PAGE..... *PolyAcrylamide Gel Electrophoresis*
- PD *Plasma Desorption*
- PMF *Peptide Mass Fingerprint*
- PRIDE..... *Protein Identifications Database*
- PSM *Peptide-Spectrum Match*
- PTM *Post-Translational Modification*

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- QIT *Quadrupole Ion Trap*
- QTOF *Quadrupole-Time Of Flight*
- QTRAP *Quadrupole-Ion Trap*
- RP-HPLC *Reverse Phase High Performance Liquid Chromatography*
- SCX *Strong Cation Exchange*
- SDS-PAGE ... *Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis*
- SLD..... *Soft Laser Desorption*
- TOF..... *Time Of Flight*
- TPP..... *Trans Proteomics Pipeline*
- TPR..... *True Positive Rate*

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

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

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

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

