Facultad de Ciencias

Departamento de Biología Molecular

DESARROLLO DE NUEVAS METODOLOGÍAS INFORMÁTICAS APLICADAS A LA ESPECTROMETRÍA DE MASAS Y AL ANÁLISIS MASIVO DE DATOS GENERADOS EN PROYECTOS DE PROTEÓMICA UTILIZANDO TÉCNICAS DE SEGUNDA GENERACIÓN

Memoria presentada para optar al grado de

Doctor en Ciencias por el licenciado

Pedro José Navarro Álvarez

Director de Tesis:

Prof. Jesús Vázquez Cobos

Madrid, 2010

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
UAM
UNIVERSIDAD AUTONOMA DE MADRID
CENTRO DE BIOLOGÍA MOLECULAR "SEVERO OCHOA"

D. Jesús María Vázquez Cobos, Profesor de Investigación del CSIC y director del trabajo de investigación titulado: "Desarrollo de nuevas metodologías informáticas aplicadas a la espectrometría de masas y al análisis masivo de datos generados en proyectos de proteómica utilizando técnicas de segunda generación", llevado a cabo por Pedro José Navarro Alvarez,

INFORMA

Que el trabajo de investigación citado ha dado lugar a resultados muy relevantes en su campo, entre los que se incluyen el desarrollo de un nuevo método (razón de probabilidades) para la identificación masiva de péptidos con prestaciones superiores a los existentes actualmente, y un nuevo método para calcular la tasa de error de identificación de péptidos usando bases de datos aleatorias que integra y mejora los dos métodos usados actualmente. Por otra parte, ha desarrollado un modelo estadístico para la cuantificación masiva de proteínas mediante marcaje isotópico estable por oxígeno-18 y análisis mediante trampa iónica lineal, que constituye el modelo más avanzado publicado hasta el momento, y, posteriormente y en la misma tesis, ha desarrollado un modelo universal para el análisis estadístico de experimentos de proteómica cuantitativa usando cualquier espectrómetro de masas o método de marcaje basado en una hipótesis nula que ha sido contrastada en todos los casos. Todo ello ha sido integrado en una plataforma software para la identificación, cuantificación y análisis estadístico de experimentos de proteómica cuantitativa, de aplicabilidad general y que integra todos los desarrollos matemáticos, algorítmicos y estadísticos realizados en esta tesis.

C/Nicolás Cabrera 1

Cantoblanco (Campus UAM)

28049-Madrid

Teléfono: +34-911964401

Fax: +34-911964420

CSIC
Consejo Superior de Investigaciones Científicas
UAM
Universidad Autónoma de Madrid
CENTRO DE BIOLOGÍA MOLECULAR "SEVERO OCHOA"

Este trabajo ha dado lugar a las siguientes publicaciones nacionales:

-Serrano H; I, J.; Martíinez-Acedo P; Navarro P; Pérez-Hernández D; Miró-Casas E; García-Dorado D; Vazquez, J. Quantitative proteomics of mitochondrial membrane proteins by sodium dodecyl sulphate polyacrylamide gel electrophoresis, 16O/18O stable isotope labeling and linear ion trap mass spectrometry. Proteómica 2007, 0, 29-34.

- Serrano, H.; Jorge, I.; P, M.-A.; P, N.; Pérez-Hernández, D.; Núñez, E.; Ramírez-Boo, M.; Bonzón, E.; Radfar, A.; Miró-Casas, E.; García-Dorado, D.; Vázquez, J.Estudio de los cambios de expresión en el proteoma de membrana mitocondrial de cardiomiocitos de rata en respuesta a un modelo de precondicionamiento isquémico. Proteómica 2008, 1, 39-40.

-Núñez E, Jorge I, Serrano H, Martínez-Acedo P, Navarro PJ, Pérez D, Miró-Casas E,García Dorado D, Vázquez J. Identificación de Conexina 32 en membranas de mitocondria de hígado y de corazón en ratones Cx43KI32

Proteómica 2008, 1, 35-36.

-Martínez-Acedo P, Martínez-Ruiz A, Maldonado AM, Horcajo-Redondo M, Jorge I, Serrano H, Navarro PJ, Pérez-Hernández D, Nuñez E, Redondo JM, Jorrín JJ, Lamas S, Vazquez J. New insights in the study of s-nitrosylation & s-nitration: strategies and problems. Proteómica 2008, 45-46.

-Navarro PJ, Jorge I, Martínez-Acedo P, Díaz M, Pérez D, Núñez E, Bonzón E, Serrano H, Vázquez J.

Desarrollo de herramientas bioinformáticas aplicadas a la identificación y cuantificación de péptidos en experimentos a gran escala. Proteómica 2008, 1, 59-60.

C/Nicolás Cabrera 1

Cantoblanco (Campus UAM)

28049-Madrid

Teléfono: +34-911964401

Fax: +34-911964420

CSIC
Consejo Superior de Investigaciones Científicas
UAM
Universidad Autónoma de Madrid
CENTRO DE BIOLOGÍA MOLECULAR "SEVERO OCHOA"

Así como a las siguientes publicaciones internacionales:

-S. Martínez Bartolomé, F. Martín-Maroto, P. Navarro, D. López-Ferrer, M. Villar, A. Ramos-Fernández, J.P. García-Ruiz, and J Vázquez. "Properties of average score distributions of SEQUEST: the probability ratio method" Molecular & Cellular Proteomics 7, 1135-1145 (2008)

-P. Navarro and J. Vázquez "An improved method to calculate False Discovery Rates for peptide identification using decoy databases" J. Proteome Res. 8; 1792–1796 (2009)

-P.Navarro/I. Jorge, P. Martínez-Acedo, E. Núñez, H. Serrano, A. Alfranca, J.M. Redondo and J. Vázquez "Statistical model to analyze quantitative proteomics data obtained by 18O/16O labeling and linear ion trap mass spectrometry:Application to the study of VEGF-induced angiogenesis in endothelial cells" Molecular & Cellular Proteomics 8; 1130-1149 (2009)

-Elena Bonzon-Kulichenko, Daniel Pérez-Hernández, Estefanía Núñez, Pablo Martínez-Acedo, Pedro Navarro, Marco Trevisán, María del Carmen Ramos, Saleta Sierra, Sara Martínez, Marisol Ruiz-Meana, Elizabeth Miró-Casas, David García-Dorado, Juan Miguel Redondo, Javier S. Burgos and Jesús Vázquez "A robust method for quantitative high-throughput analysis of proteomes by 18O labeling" (En revisión, 2009)

C/Nicolás Cabrera 1

Cantoblanco (Campus UAM)

28049-Madrid

Teléfono: +34-911964401

Fax: +34-911964420

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
UAM
UNIVERSIDAD AUTONOMA DE MADRID
CENTRO DE BIOLOGÍA MOLECULAR "SEVERO OCHOA"

Lo que me permite concluir, que, a mi juicio, el trabajo llevado a cabo por el citado licenciado tiene una calidad excelente y constituye una contribución muy relevante, a nivel internacional, en el campo de la Proteómica, habiéndose publicado parcialmente en algunas de las revistas más prestigiosas de este campo y estando en preparación al menos dos manuscritos más. Y por tanto reúne todas las cualidades necesarias para ser presentado como trabajo de tesis doctoral.

Por todo lo cual,

AUTORIZA

La defensa de la citada Tesis Doctoral, con el mismo título, por Pedro José Navarro Alvarez.

Lo que hago constar a los efectos oportunos en Cantoblanco a 26 de enero de 2010.

Fdo. Jesús Vázquez

C/Nicolás Cabrera 1

Cantoblanco (Campus UAM)

28049-Madrid

Teléfono: +34-911964401

Fax: +34-911964420

A mi padre Pedro.

¡Cuánto hubiéramos disfrutado juntos esto!

Agradecimientos

Llegar hasta aquí no debe ser más que la confirmación de haber emprendido un buen camino. El camino que, con enormes dudas asaltándome a cada paso, elegí hace algunos años recorrer. Aunque queda mucho por ver y espero que muchas más alegrías de las que disfrutar en años venideros, este es un gran momento para mí, y es preciso, para disfrutar aún más de él, recordar y agradecer a todas las personas con las que durante estos últimos años he tenido el placer de compartir experiencias.

Es difícil agradecer en pocas líneas lo que mi director de tesis, el Dr. Jesús Vázquez, ha hecho por mí en estos últimos cuatro años. Jesús transmite a todo el que se acerca a él su enorme pasión por la Ciencia, acompañada de una cercanía extraordinaria. Su paciencia, además, es casi casi ilimitada… ¡la de burradas que me has tenido que aguantar! Gracias por compensar mis eternas dudas en mí mismo con una confianza en mí prácticamente ciega. Y gracias por devolverme la pasión por mi trabajo, compartiendo tu conocimiento en innumerables y larguísimas conversaciones.

Le agradezco a mi madre Carmen su cariño, su paciencia, y que un día decidiera junto a mi padre que su sueño común fuera la educación de sus hijos, y que continuara creyendo en este sueño aún en los momentos más complicados. Gracias a ti y a papá hoy estos éxitos son también tuyos. Gracias a mi hermano David por ser mi primer modelo de científico, a mi hermana Mónica por ser mi mejor confidente y a mi hermano Miguel Ángel, por aguantar este mal carácter que a veces como buen Navarro saco.

En mi trabajo diario he tenido el enorme privilegio de contar con estupendos compañeros de fatigas. Inma y Elena: siempre habéis apoyado mi trabajo, pero desde un espíritu crítico que no se ha dejado dominar por los infinitos números, ecuaciones y jerga estadística. Gracias por vuestra amistad, y por apoyarme y criticarme al mismo tiempo: sois unas científicas maravillosas. Pablo: gracias por la amistad, por tu serenidad en el trabajo, por ser el compi perfecto de calibraciones (tú lo haces que yo te miro), por las pulguitas que hemos disfrutado, y por estar pensando ahora mismo lo que estás pensando (Peter se ha bajado los pantalones y está peloteándonos a todos). Marco: ¡Cuántos dolores de cabeza provocas! Pero los agradezco también, porque de ellos salen (alguna que otra vez) conversaciones interesantísimas y, muchas otras veces, ralladas totalmente hilarantes. Gracias por ser un amigo sincero, y por no poder ser de otro modo. Estefanía y Dani: gracias por vuestra amistad, por vuestra simpatía, y por vuestros buenos consejos. Y un agradecimiento en general a todos vosotros, por ser los grandes sufridores del “software automático”. Sin vuestras aportaciones, el resultado final no hubiera sido el mismo.

Y al otro lado del cuadrilátero (o del laboratorio), el servicio de Proteómica: Anabel, Yoli, Espe, Merche, y Sandra. Gracias por vuestra simpatía, las risas, los comentarios jocosos, y las coñas variadas. Además también te agradezco Anabel, todo lo que me has enseñado de espectrómetros de masas, secuenciación de novo, y parrilladas con la mejor carne de toda la sierra.

En el mundo de la Proteómica, pero más allá de las paredes del 311, he conocido a gente entrañable como Salva (¡qué de birras y cursos de bioinformática hemos pasado!), Benito, un

tío fenomenal que siempre tiene tiempo para explicarte mil cosas o simplemente escucharte, y Bernabé, siempre tan cordial. También he tenido algún reencuentro divertido: Ferny, después de tantos años que nos conocemos, ¿¿¿pero tú trabajabas aquí???

De los muchos que han colaborado con nosotros, me gustaría recordar a Juan Miguel Redondo y a Arantxa Alfranca, que fueron extremadamente pacientes y comprensivos esperando por sus resultados. Y, por supuesto, a todos aquellos que fuera de nuestro laboratorio han probado nuestros métodos, y nos han dado un *feedback* que ha hecho que sienta que el trabajo realizado es útil más allá del CBM: Juan Casado, Mónica Carrera, Marisa, Lola y, en especial, por todo el tiempo compartido dentro del laboratorio en sus estancias, y la amistad desarrollada, a María, Marta y Serena Camerini. Agradezco también al Dr. Jean-Charles Sánchez la oportunidad dada para dar a conocer el trabajo de nuestro grupo en su laboratorio.

De los sitios que he conocido durante la tesis, quiero destacar a Córdoba como un gran descubrimiento. Es una ciudad a la que siempre quiero volver. Agradezco a los amigos de allí por compartir conmigo vuestra ciudad, en especial a Ángela y a Jesús Jarrín y, por supuesto, a María y su familia y amigos.

Huyendo de la proteómica, quiero acordarme del primer "jefe" que tuve en un trabajo científico: el Dr. Jesús San Fabián. Gracias por el tiempo dedicado, y por dejarme compartir contigo mis inquietudes de una forma tan cercana. También, a mis amigos del máster de Biofísica: Mohammad, Sandra, Noel, Elena, Ines y Yannick. Y, por los ratos pasados durante la carrera, a toda la gente increíblemente loca que he conocido allí, en especial a los Antarianos: Jesús, los Luises (Gómez y Tortosa), Marcos (el abuelo), Josieteeeee (¡gracias por el logo del QuiXoT! Y por muchas otras cosas.), Paloma (la de tiempo que compartimos en el "cuarto oscuro" revelando fotos), Pilar y Nano, que trató de enseñarme a tocar la guitarra y tras cuatro días me dio este sabio consejo, que trato de seguir: "tú dedícate a la Ciencia, o a lo que sea, pero no te acerques nunca más a una guitarra".

A los amigos que siempre tuve a mi lado: Vero, Óscar, Ruth, Carmè y, nuevamente, Mohammad. A mis excompis de piso, que han soportado estoicamente mis cambios de humor: Raquel y Andrés, Guti, Pablo, que todo lo averigua (palabra cordobesa), y Nuria. En mi "destierro" en Guadalajara, o Villa Soledad, Guada (Guadasphere!) y Carlos me hicieron mucha compañía. Gracias a todos vosotros por los ratos que hemos pasado juntos.

En el camino también personas cercanas nos dejaron. Juan Ignacio (nuestro Chino), has sido un ejemplo de vida para mucha gente, y un orgullo para los que te conocimos.

Y en el temor de parecer excesivamente Almodovariano: quiero agradecer a la virgen de Guadalupe sus medio milagros.

Finalmente, quiero agradecer a todas aquellas personas del CBM que han hecho que sienta a este centro como mi casa.

¡Gracias a todos!

INDICE

Abstract

High-throughput identification of peptides in databases from tandem mass spectrometry data is a key technique in modern Proteomics. In this work, we introduce a novel indicator, the probability ratio, which takes optimally into account the statistical information provided by the first and second best scores obtained by the database searching engine SEQUEST. The probability ratio is a non-parametric and robust indicator that makes unnecessary spectra classification according to parameters such as charge state and allows a peptide identification performance, on the basis of false discovery rates, at least better than that obtained by other empirical statistical approaches. The indicator can also be modified to take into account the isoelectric point information obtained after IEF peptide fractionation. The probability ratio also compares favorably with statistical probability indicators obtained by the construction of single-spectrum SEQUEST score distributions. These results make the robustness, conceptual simplicity and ease of automation of the probability ratio algorithm a very attractive alternative to determine peptide identification confidences and error rates in high-throughput experiments. In the other hand, statistical models for the analysis of protein expression changes by stable isotope labeling are still poorly developed. Besides, large-scale test experiments to validate the null hypothesis are lacking. In this work we analyze several null-hypothesis, large-scale quantitative proteomics experiments performed using different isotope labeling approaches and mass spectrometry machines. Current statistical models based on normality and variance homogeneity were found unsuitable to describe the null hypothesis in all the situations tested, producing false expression changes. A random-effects model was then developed including four different sources of variance at the spectrum-fitting, scan, peptide and protein levels. With the new model the number of outliers at scan and peptide levels and the number of false expression changes were negligible in all the cases analyzed. The new model allowed to pass normality test all the three quantitation levels, becoming the first integrated, null-hypothesis tested statistical model capable of interpreting any kind of quantitative data obtained by stable isotope labeling. All these algorithms and statistical models have been integrated in a software platform called QuiXoT.

Glosario de términos utilizados

2-DE	Electroforesis bidimensional
AGC	Control de ganancia automática
CE	Energía de colisión normalizada
CID	Disociación inducida por colisión
DIGE	Electroforesis diferencial en gel
ESI	Ionización por *electrospray*
FT	Transformada de Fourier
FT-ICR	Analizador de masas de resonancia ciclotrónica de iones y transformada de Fourier (Fourier transform ion cyclotron resonance).
FTMS	Espectrometría/espectrómetro de masas por transformada de Fourier
HCD	Disociación inducida por colisión de alta energía
HPLC	Cromatografía líquida de alta presión
ICAT	Etiquetas de afinidad codificadas por isótopos
IEF	Isoeletroenfoque
IT	Tiempo de inyección (injection time)
iTRAQ	Etiquetas isobáricas para cuantificación relativa y absoluta
LIT	Trampa iónica lineal
LTQ	Cuadrupolo de confinamiento lineal
MALDI	Ionización mediante desorción por láser asistida por matriz
MS/MS	Espectrometría de masas en tándem

MudPIT	Tecnología de identificación multidimensional de proteínas
PMF	Huella de masas peptídicas
PQD	Disociación de iones mediante pulsos Q (Pulsed-Q Dissociation)
Q-TOF	Analizador de masas mediante cuadrupolo y tiempo de vuelo
RP-HPLC	Cromatografía líquida de alta presión por fase reversa
SCX	Cromatografía por intercambio catiónico fuerte
SDS-PAGE	Electroforesis en gel poliacrilamida con dodecilsulfato sódico
SILAC	Marcaje isotópico estable con aminoácidos en cultivos celulares
TMT	Etiquetas para espectrometría de masas en tándem (*Tandem Mass Tags*)

Introducción

La Proteómica moderna: estado del arte

¿Qué es la proteómica?

La proteómica es el estudio del conjunto de proteínas expresadas por un grupo de células determinadas (un tejido, un órgano, un organismo unicelular o incluso un grupo de diferentes organismos vivos que conviven en un área común) en un tiempo determinado y bajo unas condiciones específicas. Los objetivos principales de esta área son la completa descripción del conjunto de proteínas (proteómica identificativa o cualitativa), el análisis de cambios globales en los niveles de expresión de proteínas entre varias muestras (proteómica cuantitativa), el análisis de modificaciones post-traduccionales, y el estudio de interacciones proteína-proteína. Existen evidencias recientes que sugieren que el análisis cuantitativo y dinámico del proteoma completo de sistemas de relevancia biomédica podría tal vez estar pronto al alcance de la tecnología actual. Ello permitiría el análisis global no sesgado e independiente de hipótesis previas de los mecanismos moleculares de procesos fisiológicos y patológicos y también la identificación de nuevos biomarcadores con una profundidad no alcanzada anteriormente.

Proteómica Clásica o de Primera Generación.

Tradicionalmente la proteómica se ha basado en la separación de proteínas mediante electroforesis bidimensional (2DE) a partir del extracto de proteínas objeto del análisis. Este método de separación permite obtener "mapas" de proteínas que pueden ser comparados unos con otros (proteómica cuantitativa clásica). Para la identificación de cualquiera de las proteínas detectadas en estos mapas, se corta la mancha, o *spot* de la proteína que se desea identificar y se somete a un proceso de digestión mediante algúna proteasa conocida (típicamente se utiliza tripsina). Los péptidos obtenidos de esta digestión son utilizados para la identificación de las proteínas mediante espectrometría de masas (MS).

Existen dos métodos para la identificación de proteínas mediante MS. La identificación de proteínas a partir de las huellas de masas peptídicas (Peptide Mass Fingerprinting o PMF (WJ Henzel et al., 1993, P James et al., 1993, M Mann et al., 1993, DJ Pappin et al., 1993, JR Yates,

3rd et al., 1993)) se basa en la detección precisa de las masas de los péptidos de la proteína problema. Estas masas se comparan *in-sílico* con las huellas de masas peptídicas obtenidas a partir de bases de datos de proteínas. En esta técnica se utiliza típicamente un espectrómetro del tipo MALDI-TOF, que ioniza los péptidos mediante desorción por láser asistida por matriz (MALDI) y produce predominantemente iones de carga unidad. El TOF permite estimar la razón masa/carga de los iones determinando el tiempo que tardan en llegar al detector.

Otra técnica de identificación de proteínas por MS se basa en la identificación de la secuencia de los péptidos a partir de sus espectros de fragmentación (E Mortz et al., 1996) (MS^2 o MS/MS). Los espectros de fragmentación se producen al descomponer los iones de una especie química en fase gaseosa, típicamente mediante colisiones con gases neutros (collision-induced dissociation, CID). Los iones a fragmentar son previamente aislados en el interior del espectrómetro de masas, pudiéndose producir tanto fragmentos cargados como fragmentos neutros. Sólo los primeros pueden ser separados en otra etapa de MS y detectados. El espectro de fragmentación es como un "*puzzle"*, en el que las diferencias de masas entre las diferentes piezas (los iones de los fragmentos) se utilizan para secuenciar el péptido aislado. La identificación de estos espectros "puzzle" se realiza generalmente comparando el espectro obtenido con espectros teóricos generados a partir de bases de datos mediante motores de búsqueda, como se detalla en un apartado posterior. Esta técnica de identificación se lleva a cabo convencionalmente analizando los péptidos mediante nanospray en modo *off-line* y seleccionando manualmente los iones que se desean fragmentar, o en modo *on-line*, acoplando una separación peptídica extra (mediante cromatografía HPLC) previa a la ionización y el análisis en el espectrómetro de masas, que se lleva a cabo de forma automática. El método más habitual en la actualidad para identificar spots de proteínas en geles es mediante análisis en un espectrómetro MALDI-TOF-TOF, que permite obtener de forma automática tanto espectros de huella de masa peptídica de las proteínas digeridas como espectros de fragmentación de los péptidos que se detecten con mayor intensidad.

Proteómica a nivel de péptido o de Segunda Generación

En los últimos años las limitaciones de las técnicas basadas en 2DE se han hecho evidentes. Además de presentar dificultades de reproducibilidad, serios problemas en el análisis de proteínas de membrana, proteínas de gran tamaño o con un punto isoeléctrico extremo, el problema fundamental estriba en su limitado rango dinámico, que impide identificar proteínas en bajas concentraciones en presencia de otras proteínas mucho más abundantes(SP Gygi et al., 2000, SP Gygi et al., 1999). Además de esta gran limitación, debe tenerse en cuenta que esta aproximación no es adecuada para la identificación sistemática de los sitios exactos donde tienen lugar modificaciones post-traduccionales (R Aebersold,S Patterson, 1998).

En los últimos años se han desarrollado estrategias alternativas a la aproximación basada en geles; en estas técnicas (que son habitualmente denominadas como "proteómica basada en espectrometría de masas" o "proteómica de segunda generación" aunque nosotros preferimos denominarla, de manera que creemos más precisa, como "proteómica a nivel de péptido") los extractos de proteínas son digeridos en solución sin una separación previa y esta muestra compleja de péptidos es separada mediante cromatografía multidimensional acoplada al análisis mediante espectrometría de masas MS/MS (JR Yates, 3rd et al., 1995). Los espectros MS/MS son utilizados para la identificación automática de los péptidos en una base de datos. El desarrollo de esta estrategia de alto rendimiento durante la pasada década ha sido formidable. Diez años atrás, la simple secuenciación de una proteína era un logro, y actualmente la identificación de cientos de proteínas en un solo experimento es un hecho rutinario(J Cox,M Mann, 2007).

Uno de los puntos críticos de la proteómica de segunda generación es la manera como se separan los péptidos resultantes de la digestión de forma que se obtenga la máxima la profundidad del análisis. En los últimos cinco años han surgido un gran número de estrategias. Una de las primeras consistía en digerir mezclas más o menos complejas de proteínas sin separación previa. Los péptidos así producidos se prefraccionaban mediante HPLC utilizando columnas de intercambio catiónico, recolectando las fracciones que posteriormente se analizaban mediante cromatografía en fase reversa acoplada a un espectrómetro de masas en tándem (RP-HPLC MS/MS)(I Jorge et al., 2009, AJ Link et al., 1999). Otra estrategia muy utilizada es la separación de proteínas mediante SDS-PAGE monodimensional, seguida de la

digestión por separado en diversas zonas del gel y del análisis de cada fracción mediante RP-HPLC-MS/MS (GeLCMS). Estas estrategias surgen también alrededor de importantes avances en cromatografía, separación y espectrometría de masas; un ejemplo notable es la reciente aparición del analizador de masas Orbitrap, basado en el cálculo de frecuencias de oscilaciones armónicas alrededor de un electrodo de forma de huso, que puede alcanzar un poder de resolución de hasta 200,000 y una precisión de masas de 1-2 ppm. Más recientemente han aparecido técnicas de separación de los péptidos por isoelectroenfoque (IEF) en solución (OffGel(P Horth et al., 2006)). La técnica de OffGel en particular ha demostrado ser altamente reproducible y permitir un rendimiento óptimo. En un trabajo reciente, mediante una estrategia de digestión en solución, separación por IEF en OffGel y posterior análisis en un LTQ-Orbitrap se ha descrito la cuantificación del proteoma completo de levadura (4,000 proteínas)(LM de Godoy et al., 2008). En un trabajo posterior del mismo laboratorio, usando un nuevo método de solubilización con SDS y digestión "en filtro" en presencia de altas concentraciones de urea, se identificaron hasta 7,000 proteínas en células humanas en un solo experimento(JR Wisniewski et al., 2009). En nuestro laboratorio se ha puesto a punto una variante de este método, en la que se realiza la digestión en un gel SDS-PAGE concentrante en el que no se separan las proteínas, tras el proceso de digestión se separan los péptidos por OffGel, y se analizan las muestras por RP-HPLC-MS/MS en tándem con una trampa lineal LTQ(E Bonzon-Kulichenko et al., en revisión). Todas estas técnicas permiten la cuantificación relativa de dos o más muestras, como se explica en el apartado de cuantificación de proteínas por MS.

Identificación de proteínas a partir de espectros MS/MS

La fragmentación de péptidos ionizados produce diferentes tipos de iones fragmento, según el punto en que se produce la ruptura del enlace. En cada ruptura del ión parental se generan dos fragmentos, siendo detectables por MS los que permanecen cargados después de la fragmentación. Existe una nomenclatura(P Roepstorff,J Fohlman, 1984) muy extendida en proteómica que clasifica los productos de una fragmentación, el fragmento que fue ionizado y la carga de dicho ión. En la Figura I.1 se esquematizan los principios de esta anotación. Una activación del tipo CID rompe los péptidos de forma mayoritaria por el enlace peptídico, generando iones de tipo *b* e *y*. La identificación de un péptido puede realizarse de forma manual (método conocido como secuenciación *de novo*), relacionando las diferencias entre los

iones fragmento con masas de aminoácidos. Sin embargo, éste es un proceso tedioso, y en experimentos de alto rendimiento de proteómica de segunda generación, en los que deben analizarse cientos de miles de espectros MS/MS, es un método absolutamente inviable.

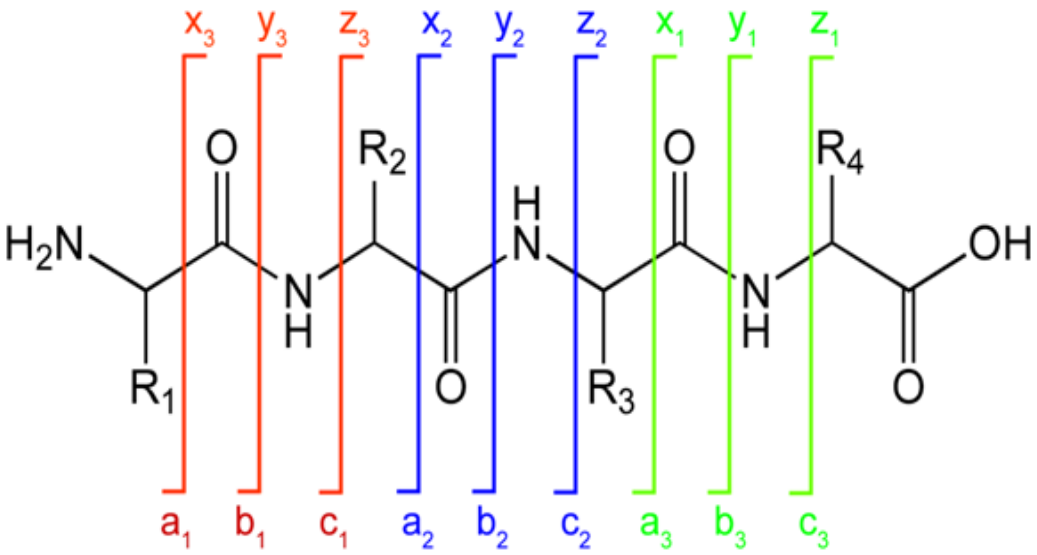


Figura I. 1 Nomenclatura según Roepstorff de la fragmentación de péptidos**. Se diferencian por colores fragmentos de iones con un número de residuos diferente. En la notación, los números indican el número de resíduos de aminoácidos contenidos en el ión al que se hace referencia, la letra (a,b,c) o (x,y,z) indica el sitio de fragmentación del ión: las letras b e y corresponden a la ruptura en el enlace peptídico, la ruptura más habitual en activación CID. Por simplicidad se muestra un péptido de sólo cuatro aminoácidos.**

Los motores de búsqueda

El método principal para la identificación de péptidos a partir de sus espectros MS/MS es la utilización de programas (llamados motores de búsqueda) que utilizan la información contenida en las bases de datos de proteínas. Estos programas realizan una digestión *in-sílico* de todas las proteínas contenidas en una base de datos y calculan las masas de los fragmentos teóricos que producirían dichos péptidos al disociarse. Los espectros MS/MS teóricos se comparan automáticamente mediante diversos algoritmos con los espectros MS/MS experimentales (véase Figura I. 2). El motor de búsqueda evalúa la semejanza entre dichos espectros calculando una cierta puntuación a esta asignación espectro – péptido (PSM, del inglés Peptide-Spectrum Match).

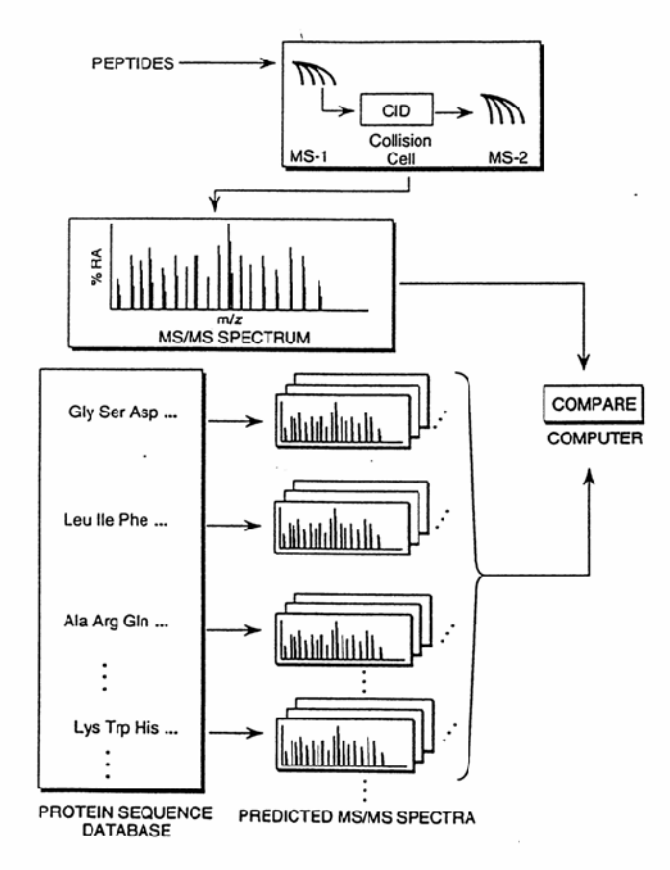


Figura I. 2 Esquema de funcionamiento de los motores de búsqueda. Los espectros de fragmentación producidos por el espectrómetro de masas son comparados uno a uno con todos los espectros teóricos generados *in-sílico* a partir de las secuencias presentes en las bases de datos de los péptidos que tienen la misma masa que el precursor y pueden ser productos de digestión de la misma enzima.

En la actualidad se dispone de numerosos motores de búsqueda, que se diferencian por el algoritmo de comparación entre los espectros real y teórico, y el tipo de puntuación. Se distinguen dos tipos de puntuación principales. Las puntuaciones basadas en factores paramétricos (como la puntuación Xcorr del motor de búsqueda SEQUEST(JR Yates, 3rd et al., 1995, JR Yates, 3rd et al., 1995)) tienen en cuenta exclusivamente características relativas a las propiedades comunes de los espectros comparados, sin tener en cuenta el resto de las PSMs.

En cambio, las puntuaciones basadas en distribuciones estiman la probabilidad de que la PSM sea correcta y por ello tienen en cuenta los resultados obtenidos con otros candidatos o por otros espectros de fragmentación. Algunos motores de búsqueda (Mascot, Phenyx, Tandem) estiman esta probabilidad utilizando todas las puntuaciones que un único espectro experimental ha obtenido al compararse con todos los espectros candidatos, formando una distribución espectral de puntuaciones que es característica de dicho espectro (*single-*

spectrum distribution). Otro método de estimación de la probabilidad de PSM correcta es seleccionar las mejores puntuaciones que obtiene cada uno de los espectros en la base de datos, y crear una distribución promedio de las mejores puntuaciones obtenidas que es característica del experimento (*average-score distribution*). El tamaño de la distribución depende del número de candidatos de cada espectro, en el primer caso, y del número de espectros MS/MS que se han obtenido en el experimento, en el segundo. Las distribuciones single-spectrum evalúan el comportamiento de cada espectro individual frente a todos los posibles candidatos en la base de datos, lo que permite determinar estadísticamente si el PSM asignado destaca significativamente con respecto a otras asignaciones posibles de ese mismo espectro con cualquier otro candidato de la base de datos. Pero en análisis de proteómica de alto rendimiento, en los que se adquieren decenas (y centenares) de miles de espectros MS/MS, es también fundamental tener en cuenta el conjunto global de mejores PSMs obtenidos en todo el experimento, siendo así posible estimar las tasas de error asociadas a la lista de péptidos identificados en dicho experimento. En la Figura I. 3 se muestra un esquema más detallado del tipo de puntuaciones utilizado por diversos motores de búsqueda.

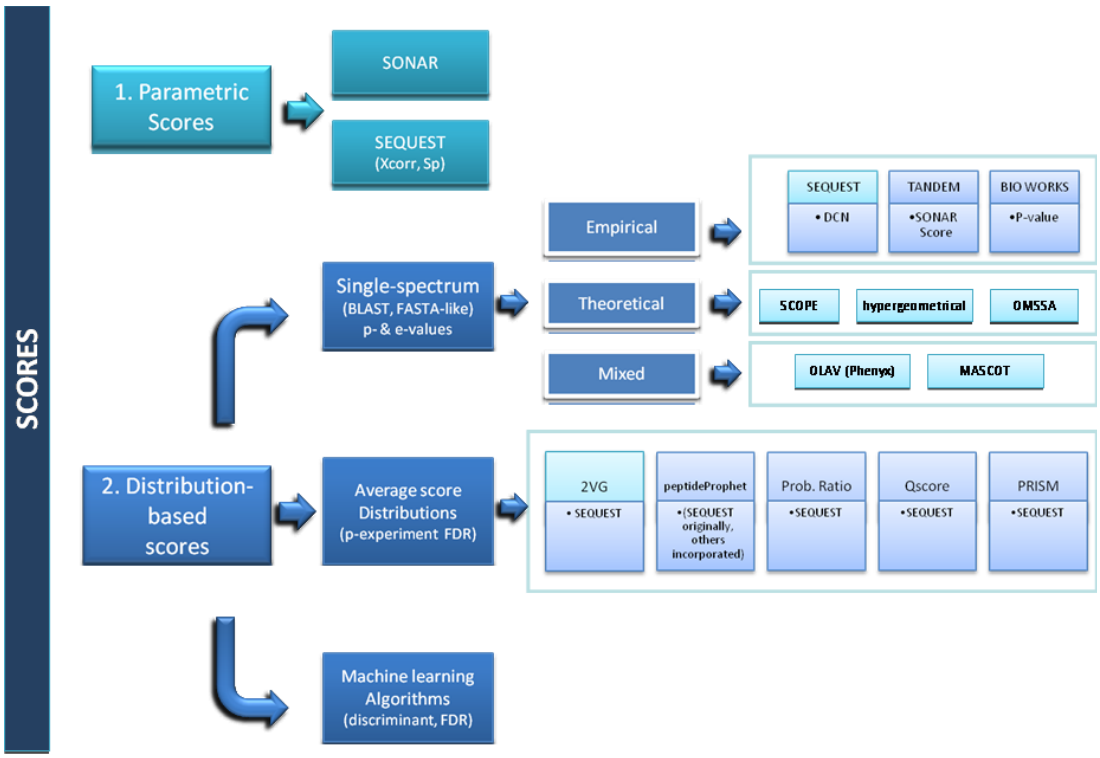


Figura I. 3 Esquema de las puntuaciones utilizadas por los motores de búsqueda más conocidos, divididos en dos grandes grupos: puntuaciones paramétricas, y puntuaciones basadas en distribuciones (puntuaciones probabilísticas).

Algoritmos estadísticos para la identificación de péptidos a partir de SEQUEST

SEQUEST fue el primer motor de búsqueda para identificar péptidos a partir de sus espectros de fragmentación, y sigue siendo uno de los más populares(AJ Link et al., 1999). Realiza una puntuación paramétrica (Xcorr) que puntúa el grado de correlación entre el espectro experimental y el teórico. Otro parámetro relacionado con Xcorr es el de la diferencia relativa de puntuación (ΔC_n), que mide la diferencia entre la mejor puntuación obtenida y la puntuación n-ésima. Los resultados de SEQUEST deben procesarse por algún tipo de algoritmo estadístico para determinar si la identificación (la que alcanza la mejor puntuación de Xcorr) puede considerarse cierta o no. Tradicionalmente se han utilizado una serie de criterios mínimos calculados de forma empírica, de manera que se suponen ciertas las asignaciones que tienen una puntuación por encima del umbral(L Florens et al., 2002, AJ Link et al., 1999, J Peng et al., 2003, WJ Qian et al., 2005, MP Washburn et al., 2001). En un trabajo previo(AJ Link et al., 1999) este método se optimizó en términos de la tasa de error (la tasa de PSMs estadísticamente incorrectas en una lista de PSMs consideradas correctas, o *False Discovery Rate*, en adelante FDR) pero posteriormente se ha demostrado que estos criterios generalmente no tienen suficiente poder discriminativo y no son suficientemente robustos para ser utilizados de forma universal en cualquier experimento (A Keller et al., 2002, RG Sadygov,JR Yates, 3rd, 2003, DL Tabb et al., 2003). Se han desarrollado criterios alternativos basados en el análisis estadístico de la distribución de puntuaciones de SEQUEST para mejorar el discernimiento de PSMs correctas e incorrectas (DC Anderson et al., 2003, A Keller et al., 2002, T Kislinger et al., 2003, D Lopez-Ferrer et al., 2004, MJ MacCoss et al., 2002, RE Moore et al., 2002, J Razumovskaya et al., 2004). PeptideProphet (A Keller et al., 2002), uno de los algoritmos más usados hoy en día, utiliza un grupo de experimentos de entrenamiento para estimar qué parámetros de SEQUEST distinguen de la mejor manera posible las PSMs correctas de las incorrectas, y determina un único parámetro de calidad mediante una función discriminante en forma de combinación lineal de dichos parámetros. Aunque se han obtenido excelentes resultados usando PeptideProphet en numerosos estudios, los resultados obtenidos dependen del grupo de experimentos elegidos para su entrenamiento, por lo que no es universalmente aplicable a cualquier experimento nuevo y, además no se ha demostrado que la combinación lineal de parámetros que utiliza sea el procedimiento óptimo para la discrimación entre espectros verdaderos y falsos. Un método desarrollado en nuestro laboratorio, denominado 2VG, utiliza la distribución de las puntuaciones Xcorr y ΔC_n

combinadas de forma empírica de forma que se ajustan a una gaussiana bidimensional; aunque los resultados obtenidos con el método 2VG son comparables e incluso superiores a los de PeptideProphet en términos de rendimiento, este método requiere el ajuste preciso de valores paramétricos que cambian de un experimento a otro, lo que dificulta la automatización(D Lopez-Ferrer et al., 2004).

Recientemente hemos desarrollado en nuestro laboratorio un método basado en la razón de probabilidades (*probability Ratio*, o pR), que relaciona las distribuciones promedio de la mejor puntuación y de la segunda mejor puntuación. El método se basa en un análisis teórico llevado a cabo por Fernando Martín-Maroto en el que se expresan las distribuciones promedio en forma de combinaciones lineales de distribuciones espectrales individuales y a partir del cual se deduce analíticamente cuál es la probabilidad de cada espectro obtenga al azar una mejor puntuación en la base de datos, cuando ha obtenido una segunda mejor puntuación aleatoria. Cuando el número de candidatos presentes en la base de datos es suficientemente alto, el análisis teórico demuestra que esta probabilidad se reduce al cociente entre los valores de las distribuciones promedio de la primera y segunda mejores puntuaciones. Este método une a su sencillez conceptual, facilidad de cómputo y ausencia de funciones o parámetros ajustables la ventaja de que, suponiendo que el modelo fuera correcto (lo que está apoyado por toda la información obtenida hasta el momento), sería el indicador más preciso que pudiera obtenerse a partir de la primera y segunda mejores puntuaciones. El procedimiento de cómputo de este parámetro se describe en más detalle en la sección de Material y Métodos.

Este nuevo método se probó con un algoritmo preliminar, mostrando unos resultados prometedores que sugerían que la pR podría mejorar a otros algoritmos. Sin embargo, en el momento de comenzar esta tesis el algoritmo todavía no estaba completamente automatizado, su comparación con otros algoritmos era todavía incompleta y era necesario corregir posibles efectos numéricos en el cálculo de la razón de probabilidades antes de poder aplicar este método en la práctica. Por otra parte, en el modelo teórico desarrollado por Martín-Maroto no se contemplaba el uso de la información que suministran las separaciones por IEF sobre el punto isoeléctrico estimado para cada secuencia identificada; esta información podría resultar muy útil para mejorar el rendimiento de las identificaciones.

El problema de la estimación de la Tasa de Error (FDR)

La identificación a gran escala de péptidos a partir de sus espectros de fragmentación es un ejemplo clásico de aplicación a la práctica de una hipótesis múltiple, donde considerar que cada asignación péptido-espectro es o no correcta puede considerarse una hipótesis independiente. En este tipo de problemas uno de los mejores métodos para controlar la eficiencia de la inferencia estadística es la tasa de error, o *False Discovery Rate* según la definición clásica (JD Storey,R Tibshirani, 2003). La tasa de error es la proporción de asignaciones incorrectas que se espera que haya en la lista de asignaciones correctas en un experimento dado. El número de PSMs incorrectas puede estimarse realizando la búsqueda contra bases de datos *aleatorias o señuelo* del mismo tamaño que la base de datos original (o base de datos *objetivo*). La búsqueda de los espectros MS/MS de un experimento de proteómica contra una base de datos aleatoria ofrece una distribución de puntuaciones aleatorias que explican cómo se comportan las puntuaciones de PSMs mal asignadas en ese experimento concreto y con el motor de búsqueda utilizado(DL Tabb, 2008).

Dada una cierta distribución de puntuaciones en un experimento, la separación entre PSM correcta o incorrectamente asignadas se realiza escogiendo un cierto valor umbral de puntuación. La forma más sencilla de calcular la FDR asociada al conjunto de PSMs con una puntuación mejor que la puntuación umbral es determinar la distribución de las puntuaciones obtenidas en una búsqueda en una base de datos señuelo (en la que todas las PSMs están mal asignadas por definición), y contar el número de PSMs que se encuentran por encima del umbral establecido, que constituye una estimación del número de falsos cometidos en el conjunto de PSMs seleccionado en la búsqueda contra la base de datos objetivo. Aunque este método se ha utilizado con mucha frecuencia en la bibliografía, esconde dos problemas principales: en primer lugar, el conjunto de PSMs de la búsqueda objetivo incluye tanto asignaciones correctas como incorrectas, por lo que el número de PSMs incorrectas estimado por este método está sobreestimado. Y en segundo lugar, los espectros MS/MS que corresponden a PSMs correctas tienden a ofrecer puntuaciones mejores en las búsquedas en bases de datos señuelo que las PSMs incorrectas, con lo que la distribución de puntuaciones en la búsqueda señuelo no refleja de forma precisa el comportamiento de las asignaciones falsas.

La forma más divulgada de solucionar estos problemas es establecer una estrategia de competición por los espectros MS/MS(JE Elias,SP Gygi, 2007). Este método se basa en la idea de que dada una PSM correcta, la puntuación de la PSM de la base de datos objetivo debe ser siempre superior que la PSM correspondiente al mismo espectro MS/MS en la base de datos señuelo. Por lo que si se utiliza una base de datos compuesta por las bases de datos objetivo y señuelo, el número de PSMs falsos se puede estimar de forma precisa multiplicando por dos el número de PSMs que pertenecen a la base de datos señuelo y que se encuentran por encima del umbral deseado. Este método, no obstante, tampoco se considera perfecto, porque la FDR no se calcula en la población original (PSM identificadas en la base de datos objetivo), sino en una población artificialmente "hinchada" conteniendo secuencias peptídicas aleatorias. La existencia de dos estrategias diferentes ha complicado en los últimos años la comparación de resultados y la interpretación de las tasas de error obtenidas en experimentos de análisis masivo. En este contexto, cabría cuestionarse si podría haber alguna manera de aunar las propiedades de ambas estrategias (bases de datos separadas y bases de datos concatenadas) en un método integrado, utilizando correctamente toda la información que ofrecen las dos estrategias.

Cuantificación de proteínas por MS

Proteómica cuantitativa basada en espectrometría de masas

En proteómica clásica, la cuantificación relativa se realiza utilizando electroforesis 2D y teñido de proteínas (J Fievet et al., 2004, GB Smejkal et al., 2004) o un pre-marcaje isotópico (Y Hu et al., 2003, JX Yan et al., 2002). Los *spot* de proteínas que muestran diferencias cuantitativas son procesados por espectrometría de masas para su identificación (A Shevchenko et al., 1996). En las aproximaciones de proteómica de segunda generación, la cuantificación no se realiza directamente sobre las proteínas, sino que éstas son digeridas previamente, por lo que la cuantificación debe realizarse en cada péptido producto de la digestión. Dicha cuantificación puede realizarse mediante espectrometría de masas, de forma relativa y utilizando marcajes isotópicos que muestren una diferencia de masas en los espectros MS o MS/MS de cada péptido analizado, o también de forma absoluta, utilizando péptidos marcados isotópicamente como estándares internos (M Bantscheff et al., 2007).

Cuantificación relativa mediante marcaje isotópico

En los últimos diez años han surgido numerosos métodos de marcaje basados en isótopos estables. Los marcajes pueden realizarse mediante reacción química con un reactivo de grupo marcado isobáricamente (ICAT (SP Gygi et al., 1999), iTRAQ (PL Ross et al., 2004) y TMT (A Thompson et al., 2003)), mediante la incorporación metabólica de aminoácidos marcados (SILAC (SE Ong et al., 2002)) o mediante marcaje enzimático con ^{18}O usando tripsina o proteasas análogas (OA Mirgorodskaya et al., 2000, A Ramos-Fernandez et al., 2007, X Yao et al., 2001).

La cuantificación relativa en espectrometría de masas se realiza relacionando las intensidades de pares de picos correspondientes a péptidos iguales, que se encuentran separados por la diferencia de masa que corresponde al marcaje isotópico. En el caso de los marcajes ICAT, SILAC y ^{18}O, los péptidos se cuantifican sobre un espectro MS, en el que el péptido se presenta como un doblete de envolturas isotópicas, y se cuantifica la diferencia

relativa de las intensidades de estas envolturas isotópicas. ICAT fue uno de los primeros marcajes isotópicos desarrollados (SP Gygi et al., 1999). Consta de dos compuestos químicos equivalentes (uno con isótopos ligeros y otro con isótopos pesados) que modifican los residuos de cisteínas y están unidos a una molécula de biotina. En sus inicios, el isótopo pesado estaba marcado con deuterio, pero este elemento químico producía retrasos en la elución cromatográfica, por lo que fue sustituido por ^{13}C y aplicado con éxito en estudios proteómicos (DK Han et al., 2001). En el método ICAT los péptidos modificados en Cys se purifican mediante cromatografía de afinidad usando columnas de avidina, de manera que sólo se cuantifican los péptidos que contienen algún residuo de Cys. Esta particularidad, por un lado, simplifica el proteoma en estudio, permitiendo alcanzar una mayor profundidad en términos del número de proteínas cuantificadas; sin embargo, por otra parte, esta simplificación del proteoma disminuye la cobertura de secuencia de cada una de las proteínas, lo que puede redundar en una menor precisión del estudio en experimentos de alto rendimiento. El marcaje con SILAC, en cambio, es un marcaje en el que se introducen aminoácidos marcados con isótopos pesados (^{2}H, ^{13}C o ^{15}N) en un cultivo celular. Esto ofrece dos ventajas importantes: la primera de ellas es que una elección conveniente de los aminoácidos de marcaje asegura que la práctica totalidad de los péptidos se encuentren marcados (una práctica muy habitual en proteómica es utilizar Lys y Arg marcadas, ya que al utilizar tripsina como enzima de corte se asegura que prácticamente todos los péptidos tienen su extremo carboxilo marcado). La segunda ventaja es que el marcaje se realiza durante el crecimiento del cultivo celular, lo que reduce la variabilidad analítica ya que las muestras que se comparan pueden ser mezcladas antes de realizar el lisado de las células. Sin embargo, esta técnica sólo es aplicable a muestras que puedan ser cultivadas en presencia de los compuestos marcados isotópicamente y, por tanto, no puede ser usada con muestras de tejidos o fluidos humanos como suero o plasma. El marcaje isotópico con ^{18}O también permite marcar prácticamente todos los péptidos de la muestra, tras la digestión tríptica del extracto de proteínas, ya que marca todos los péptidos que tengan una Lys o Arg en su extremo carboxilo. Aunque esta última técnica es de aplicabilidad universal y es económicamente la más ventajosa, el marcaje con ^{18}O se considera una técnica delicada y que requiere un riguroso control de la eficiencia de marcaje (véase el apartado siguiente).

La cuantificación mediante marcaje con iTRAQ no se realiza sobre espectros MS, sino sobre los espectros MS/MS. Existen cuatro posibles reactivos marcadores (recientemente se han ampliado a ocho) que permiten comparar hasta cuatro muestras en el mismo experimento. Están formados por un grupo reactivo que se une a grupos amino, un grupo reportero y un

grupo de balance de masas. Los cuatro reactivos son isobáricamente iguales, por lo que los péptidos marcados con cualquiera de estos reactivos tienen el mismo peso molecular y por tanto se fragmentan a la vez. Durante la fragmentación, el enlace entre el grupo reportero y el grupo balance de masas se rompe fácilmente, dando lugar en el espectro de masas a iones correspondientes a las masas de los cuatro grupos reporteros (distintas entre sí) (véase Figura I. 4). La intensidad de dichos iones permite la cuantificación relativa de las cuatro especies

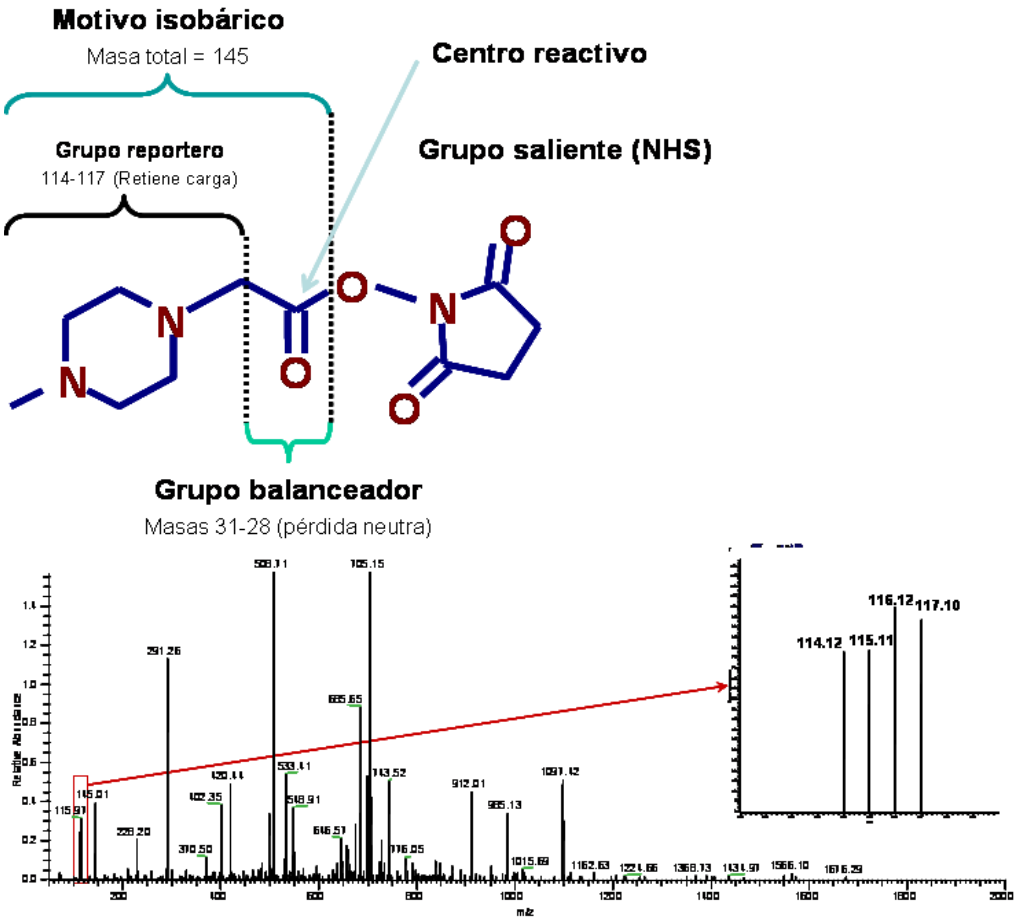


Figura I. 4 Reactivos iTRAQ. Estos compuestos contienen un grupo reactivo (éster de N-hidroxisuccinimida), que reacciona selectivamente con grupos amino (N-terminal y Lys), y un motivo isobárico formado por un grupo reportero, cuya masa está en el rango de 114 a 117 (en iTRAQ 4-plex), y un grupo de balance, cuya masa equilibra la masa del reportero de forma que el motivo isobárico siempre tiene una masa total de 145 Da. El enlace entre el grupo reportero y el grupo de balance tiene una alta eficiencia de fragmentación, por lo que en los espectros MS/MS se detectan los grupos reporteros por separado, permitiendo así la cuantificación específica de las cuatro especies por separado.

marcadas.

Marcaje isotópico con ^{18}O y el problema de la eficiencia de marcaje

El marcaje con ^{18}O es un marcaje enzimático realizado a nivel de péptido, en el que mediante la acción del enzima tripsina se marca el grupo carboxílico de los péptidos, sustituyendo en este grupo sus dos ^{16}O por dos átomos de ^{18}O. Se ha considerado clásicamente un método delicado y poco reproducible: la eficiencia del marcaje debe ser estrictamente controlada, ya que en un marcaje enzimático no todos los péptidos se marcan con la misma facilidad. La exigua distancia de marcaje (4 Da para el intercambio completo de los dos oxígenos del extremo carboxilo, y sólo 2 Da para un marcaje incompleto), por otro lado, causa un solapamiento en las envolturas isotópicas de las especies no marcada, marcada de forma incompleta, y marcada completamente. Este solapamiento dificulta el cálculo de intensidades de cada una de las especies involucradas. Además, diversos factores pueden causar un desmarcado; valores extremos de pH causan un intercambio de oxígeno con el medio (M Schnolzer et al., 1996), por lo que este tipo de marcaje no es compatible con todos los métodos de fraccionamiento peptídico. En nuestro laboratorio se han desarrollado recientemente dos aproximaciones que facilitan que el método de marcaje con ^{18}O pueda utilizarse en análisis de alto rendimiento de forma sencilla y automatizada. Por un lado, se ha desarrollado un algoritmo de cálculo de la eficiencia de marcaje para cada péptido analizado. El algoritmo tiene en cuenta la existencia de cuatro especies coexistentes: la especie procedente de la muestra no marcada (especie A), la especie marcada con dos ^{18}O (especie B_2), la marcada con un único ^{18}O (especie B_1) , y la especie que corresponde a la muestra marcada, pero sin ningún ^{18}O (especie B_0). En la Figura I. 5 se muestra un esquema de la contribución de cada una de estas especies en un espectro real. Las tres especies que corresponden a la especie marcada son relacionadas entre sí utilizando un modelo cinético del que se obtiene un único factor que describe el avance de la reacción de marcado, o *eficiencia de marcaje*. El control riguroso de este parámetro permite llevar a cabo experimentos de cuantificación diferencial sin introducir artefactos debidos al marcaje (A Ramos-Fernandez et al., 2007). Más recientemente se ha desarrollado en el laboratorio un minucioso protocolo de marcaje con ^{18}O, separación mediante OffGel, y análisis mediante HPLC-LIT-MS (E Bonzon-Kulichenko et al., en revisión). Este protocolo, que controla perfectamente que el marcaje sea homogéneo y constante en toda la muestra, y demuestra que la separación OffGel es compatible con el

marcaje ^{18}O en cualquier rango de pH, utiliza SDS como disolvente de las proteínas, siendo por tanto de aplicabilidad prácticamente universal.

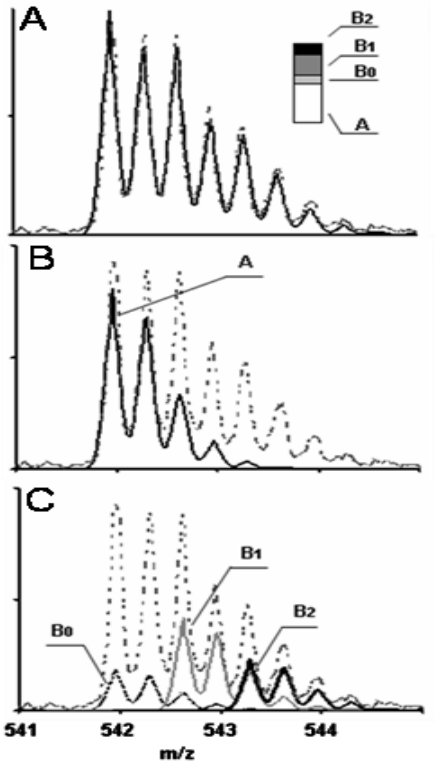


Figura I. 5 Estimación de la eficiencia de marcaje con ^{18}O. Se señala la contribución a la envoltura isotópica de cada una de las especies del péptido (A: especie proveniente de la muestra no marcada; B_0, B_1 y B_2 : especies provenientes de la muestra sometida a marcaje en la que se han incorporado ninguno, uno o dos átomos de ^{18}O, respectivamente). El algoritmo descompone la envoltura isotópica total en la suma de los cuatro componentes, que se relacionan entre sí por la eficiencia de marcaje.

Espectrometría de masas usada para cuantificación relativa de péptidos mediante marcaje isotópico estable

Los problemas de resolución son una preocupación común en la proteómica cuantitativa basada en MS y marcaje isotópico, por lo que tradicionalmente se han utilizado equipos de media y alta resolución, como los analizadores de masas FT-ICR (WJ Qian et al., 2005, X Yao et al., 2003, X Yao et al., 2001) (basado en la determinación de la frecuencia ciclotrón mediante transformadas de Fourier) o los analizadores de tiempo de vuelo (TOF). Estos equipos siguen siendo los preferidos para este tipo de análisis. Tradicionalmente se ha considerado que los equipos de baja resolución no tienen poder resolutivo suficiente para realizar proteómica cuantitativa(DS Kirkpatrick et al., 2005). La aparición reciente de la moderna trampa lineal (LIT), cuyo equipo más característico es el modelo LTQ, que se basa en el confinamiento de iones en una región vacía o tubo mediante una combinación de campos eléctricos constantes y voltajes de radiofrecuencia, ha cambiado el panorama. Estos equipos tienen una velocidad de barrido, una sensibilidad y una capacidad de iones muy superiores a los equipos de trampa iónica 3-D convencionales (V Mayya et al., 2005); las prestaciones del modelo LTQ de trampa lineal en el análisis rápido y versátil de péptidos en mezclas complejas ha sido demostrada en nuestro laboratorio(I Jorge et al., 2007). Los LIT permiten modos de barrido de media resolución sobre un rango de masas limitado (ZoomScan) que no muestran los efectos espacio-carga de las trampas 3-D; por ello se han convertido en equipos muy atractivos para realizar proteómica cuantitativa. Para ello, se pueden programar ciclos de barrido en el instrumento que realicen dos espectros, un ZoomScan centrado alrededor del doblete de envolturas isotópicas que se desea cuantificar y un espectro MS/MS para identificar el péptido; la aplicación práctica de este tipo de barrido en la cuantificación a gran escala de péptidos marcados con ^{18}O mediante LIT ha sido recientemente demostrada en nuestro laboratorio (I Jorge et al., 2009, D Lopez-Ferrer et al., 2006, A Ramos-Fernandez et al., 2007).

El marcaje mediante iTRAQ exige una alta sensibilidad en un rango de masas pequeñas, ya que los iones reporteros tienen una masa desde 114 Da a 117 Da, por lo que se realiza de manera óptima en equipos con fragmentación cuadrupolar como el Q-TOF o el TOF-TOF. Hasta hace menos de dos años no se podían utilizar trampas iónicas con esta tecnología, ya que la fragmentación por CID sigue una regla particular, conocida como la *regla del tercio*, por la que los fragmentos cuya razón masa/carga es inferior a un tercio de la del ión aislado no pueden ser confinados en la trampa en las condiciones de radiofrecuencia que permite la

fragmentación eficaz de los péptidos. Este problema se ha resuelto con una nueva tecnología de activación de disociación inducida por colisión con pulsos Q (PQD), que cambia muy rápidamente los voltajes de confinamiento durante el lapso de tiempo que los iones permanecen excitados por la colisión antes de fragmentarse. La tecnología PQD permite acceder a las trampas lineales a la zona baja del espectro de fragmentación, si bien a costa de una concomitante pérdida de sensibilidad. En los equipos LTQ-Orbitrap pueden llevarse a cabo análisis cuantitativos del tipo iTRAQ, bien con activación PQD en la trampa lineal, o bien fragmentando por HCD (JV Olsen et al., 2007) fuera de la trampa lineal y adquiriendo el espectro de fragmentación en la cámara Orbitrap.

Otra cuestión importante es la calidad de las cuantificaciones de cada péptido, que depende de forma crítica de la metodología utilizada. Generalmente no todos los péptidos identificados pueden cuantificarse con la misma precisión(LM de Godoy et al., 2008, I Jorge et al., 2009), por lo que se limita en gran manera la cobertura del proteoma. El marcaje por SILAC y por ^{18}O son métodos que cuantifican a nivel de espectros MS, mientras que iTRAQ utiliza la información obtenida a nivel MS^2, que en teoría debería ofrecer un mayor rango dinámico ya que los espectros MS^2 generalmente tienen un ruido de fondo mucho menor (M Bantscheff et al., 2008). Sin embargo, la contribución de especies coeluyentes y cercanas isobáricamente puede contribuir a la intensidad de los iones reporteros en iTRAQ, disminuyendo la precisión de las determinaciones de cambios de expresión (M Bantscheff et al., 2008). Por supuesto, en tecnologías de marcaje cuantificadas sobre espectros MS (^{18}O, SILAC) las especies coeluyentes también pueden afectar a las cuantificaciones, pero a diferencia de iTRAQ, su presencia pueden ser detectadas en el espectro MS (en baja resolución), o no afecta al no contribuir sobre ninguno de los picos isotópicos del péptido analizado (en alta resolución). Este efecto es previsiblemente más problemático bajo condiciones de alta concentración de péptidos, que son las que habitualmente se dan en experimentos de alto rendimiento.

Análisis estadístico de experimentos de cuantificación diferencial por marcaje isotópico a gran escala.

El número de datos que deben estudiarse en experimentos de proteómica diferencial a gran escala es formidable. Esto hace que uno de los problemas principales de la proteómica cuantitativa a gran escala sea el análisis de datos. Particularmente uno de los cuellos de botella fundamentales es determinar si los cambios de expresión de proteínas observados son correctos, o se han producido de forma artifactual. Los cambios de expresión en proteínas pueden considerarse significativos si son cambios mayores que un cierto umbral fijo que se establezca a priori, pero es preferible utilizar métodos estadísticos que establezcan la significatividad estadística basándose en criterios de probabilidad. En este sentido, los métodos estadísticos aplicados en proteómica cuantitativa están claramente menos desarrollados que los que se utilizan en el campo de micro matrices. En este campo se ha realizado ya una detallada discusión sobre el problema de la hipótesis múltiple, el control de la tasa de error (FDR)(SS Li et al., 2005) y el uso del valor q (JD Storey,R Tibshirani, 2003) como parámetro de inferencia. Por ejemplo, existe actualmente una controversia muy grande en torno a la validez de la suposición de que los valores p generados por los métodos estadísticos habitualmente utilizados en micro matrices siguen una distribución normal (AA Fodor et al., 2007). En cambio, los trabajos sobre métodos estadísticos aplicados a datos de proteómica cuantitativa son muy escasos. La mayoría de las publicaciones de proteómica cuantitativa mediante MS han heredado los métodos ya establecidos en geles de electroforesis bidimensional, sin comprobar si estos métodos son realmente aplicables a las aproximaciones de MS/MS. Sólo recientemente la comunidad proteómica ha tomado conciencia de la necesidad de utilizar diseños experimentales robustos (NA Karp,KS Lilley, 2007) y de controlar la FDR en experimentos de electroforesis cuantitativa (NA Karp et al., 2007). Sin embargo, todos estos tests asumen un muestreo independiente y que los datos analizados siguen una distribución normal, suposiciones que no han sido demostradas en un contexto general. La situación en proteómica cuantitativa basada en isótopos estables y posterior análisis por MS es muy parecida. Se han desarrollado numerosos métodos bioinformáticos que permiten el análisis automatizado de datos de MS, y el cálculo de las proporciones de los péptidos observadas para cada una de las estrategias de marcaje isotópico (KW Lau et al., 2007, LN Mueller et al., 2008), pero siguen siendo escasos los trabajos relacionados con la determinación estadística de cambios de expresión significativos. Prácticamente todos los métodos existentes tienen en común la suposición subyacente de que los datos de hipótesis

nula siguen una distribución normal, ya que utilizan por lo general tests de la t de Student (MJ MacCoss et al., 2003, G Wang et al., 2006) o se calculan los valores p de cada cuantificación (XJ Li et al., 2003). Sin embargo, la validez de esta suposición no ha sido todavía analizada en detalle. Sólo algunos estudios recientes han empezado a estudiar la fiabilidad y las fuentes de error en análisis de iTRAQ (AM Boehm et al., 2007, CS Gan et al., 2007), y se ha desarrollado un algoritmo de máxima verosimilitud que estima los intervalos de confianza de las proporciones de abundancia de proteína determinados por marcaje metabólico $^{14}N/^{15}N$ (V Lanquar et al., 2007, CJ Nelson et al., 2007). Siguen faltando estudios de prueba que traten de establecer las distribuciones de hipótesis nula, parecidos a los que ya se han realizado en datos de micro matrices (AA Fodor et al., 2007), y desarrollos de métodos estadísticos de aplicabilidad universal en cualquier marcaje isotópico.

Objetivos

El gran aumento en el tamaño de los datos adquiridos en experimentos de proteómica y las limitaciones de los métodos actuales hacen muy necesario el desarrollo de nuevos algoritmos y herramientas de análisis estadístico que permitan a la comunidad científica analizar de un modo robusto y fiable experimentos de proteómica identificativa y cuantitativa adquiridos mediante MS. De acuerdo al estado actual de la técnica y los puntos tratados en la Introducción, en esta tesis nos planteamos los siguientes objetivos:

1. Mejorar la inferencia estadística de identificaciones válidas de espectros MS^2 obtenidas en experimentos de alto rendimiento, basándose en el modelo teórico de la razón de probabilidades.

2. Desarrollar un modelo estadístico para el análisis de datos de expresión diferencial de proteínas mediante marcaje isotópico con $^{16}O/^{18}O$ y espectrometría de masas de trampa iónica lineal.

3. Generalizar el modelo anterior hacia un modelo estadístico universal válido para cualquier tipo de marcaje isotópico y espectrometría de masas.

4. Desarrollar un software que permita la identificación, cuantificación y análisis estadístico de datos de proteómica de expresión diferencial en experimentos de alto rendimiento.

Material y Métodos

Este trabajo está enfocado en el marco de desarrollo de métodos estadísticos y de software específico de proteómica identificativa y cuantitativa. Ya que su propósito no es el análisis de una muestra concreta, sino el desarrollo de la tecnología a aplicar para el análisis de muestras, la preparación de las mismas será comentada de forma concisa, indicando solamente lo estrictamente necesario para seguir el desarrollo experimental utilizado a lo largo del trabajo, dando una mayor importancia a los métodos matemáticos y desarrollos de software utilizados.

Preparación de muestras

Muestras utilizadas para la optimización de la validación de resultados obtenidos mediante motores de búsqueda

Para el desarrollo de métodos para experimentos de proteómica de identificación de proteínas a gran escala se empleó como proteoma modelo un extracto de proteínas de núcleo de células tipo T Jurkat, obtenidos según se describe previamente(AL Armesilla et al., 1999). Brevemente, un extracto de 100 µg de proteínas se precipitó con acetona y se liofilizaron a sequedad. Los puentes disulfuro de los residuos de cisteínas fueron reducidos con DTT 10 mM durante una hora en tampón bicarbonato de amonio 25 mM, (pH 8) conteniendo urea 8 M, y seguidamente alquilados con iodo acetamida 50 mM durante 45 minutos en condiciones de oscuridad. La mezcla se diluyó 4 veces para reducir la concentración de urea y se sometió a digestión con tripsina (Promega) a 37 ºC durante toda la noche empleando una relación enzima:sustrato de 1:50. Los péptidos resultantes se separaron por cromatografía de intercambio catiónico seguida por una separación *on-line* por fase reversa acoplada a una trampa iónica lineal LCQ-DECA XP (Thermo Finnigan). Para el intercambio catiónico se emplearon columnas 0.18x150 mm BioBasic SCX (ThermoHypersil-Keystone) en un sistema microHPLC Smart (Pharmacia) y los péptidos fueron eluidos con un gradiente de KCl de 0 a 122 mM de en tampón fosfato 5 mM, (pH 3) en presencia de 25 % acetonitrilo al 25%. (ACN). Para la fase reversa se emplearon columnas 0.18x150 mm BioBasic C-18 RP (ThermoHypersil-Keystone) conectadas a un sistema HPLC Surveyor (Thermo Finnigan), eluyendo los péptidos con un gradiente de 0 a 48 % de ACN en presencia de 0.5 % de ácido acético. El espectrómetro

de masas se programó para realizar a lo largo de todo el gradiente cromatográfico la fragmentación MS^2 de los 3 iones más intensos de un barrido desde 400 hasta 1600 amu (8 µscans, 200 ms IT, 10000 AGC), empleando exclusión dinámica.

Muestras utilizadas para el desarrollo de un modelo estadístico para la cuantificación de péptidos mediante diferentes marcajes isotópicos.

En el desarrollo de métodos de proteómica cuantitativa se emplearon varios proteomas modelo. En el caso del marcaje con ^{18}O se emplearon extractos proteicos de un cultivo de células HUVEC estimuladas con VEGF 50 ng/ml de durante 0, 2 y 4 horas, obtenidos según se describe previamente (I Jorge et al., 2009). Las muestras fueron reducidas, alquiladas y digeridas con tripsina en solución tal y como se describe en el apartado anterior. Los péptidos resultantes se desalaron en cartuchos de C18 OASIS (Waters), se secaron en evaporador centrífugo y se guardaron a -20 ºC hasta ser utilizados para el marcaje diferencial con ^{18}O.

Para la generalización del modelo estadístico a otros marcajes isotópicos y otros espectrómetros de masas se emplearon extractos de proteínas de un cultivo de *Saccharomyces cerevisiae* tratado con H_2O_2 0,5 mM (muestra B) o sin tratar (muestra A). Para el marcaje con SILAC se obtuvo una tercera muestra (A*), en la cual las levaduras se cultivaron en un medio donde la totalidad de los residuos de Arg y Lys estaban marcados con ^{13}C. Los extractos de proteínas de *Saccharomyces cerevisiae de la cepa* YMJ38 se obtuvieron según se describe previamente (ML Hernaez et al., 1998). Las muestras a comparar se sometieron a electroforesis en un gel de SDS-PAGE hasta que todo el proteoma quedó concentrado en una única banda en la interfase entre el gel concentrador y el gel separador. Los geles se tiñeron con Azul de Coomasie Brilliant Blue R-250 (Bio-Rad). Se cortó la banda correspondiente a cada proteoma y se sometió a digestión con tripsina "en gel" en una relación enzima:sustrato de 1:5, en bicarbonato amónico 50 mM, (pH 8.8) (A Shevchenko et al., 2006), en presencia de 10 % de acetonitrilo y 0.01 % (p/v) 5-ciclohexil-1-pentil-ß-D-maltósido (CYMAL-5). Se extrajeron los péptidos resultantes, se desalaron en cartuchos de C18 OASIS, se secaron en evaporador centrífugo y se guardaron a -20 ºC hasta ser utilizados para el marcaje diferencial con $^{16}O/^{18}O$ ó iTRAQ, o directamente para su análisis por MS en el caso de muestras para SILAC.

El marcaje diferencial de los péptidos con $H_2{}^{16}O$ u $H_2{}^{18}O$ (95 %, Isotec) se realizó en acetato amónico 100 mM, pH 6, conteniendo un 20% (v/v) de ACN, en una relación tripsina inmovilizada: sustrato de 1:200(w/v) (p:v) durante toda la noche a 37 ºC. La reacción de marcaje se detuvo mediante la eliminación física de la tripsina inmovilizada por filtración (Wizard minicolumns, Promega) y la inhibición de la posible tripsina soluble remanente añadiendo TLCK a una concentración final de 1 mM. Seguidamente, los péptidos se mezclaron, se desalaron en cartuchos C18 OASIS usando como solución de elución 50% (v/v) de ACN en formiato amónico 5 mM, pH 3, y se separaron en 24 fracciones mediante isoelectroenfoque en un rango de pH de 3 a 10, usando un equipo OFFGEL 3100 (Agilent). Las fracciones se desalaron usando puntas OMIX C18 (Varian) y se analizaron por RP-HPLC acoplado a una trampa iónica lineal tipo LTQ o a un LTQ-Orbitrap, ambos de Thermo Finnigan. La trampa iónica lineal LTQ se programó para realizar a lo largo de todo el gradiente cromatográfico un ZoomScan (10 µscans, 100 ms IT, 3 000 AGC, ventana de -6 y +8 amu), y una fragmentación en modo CID (3 µscans, 100 ms IT, 10000 AGC, ventana de aislamiento 2 amu, 35% CE, 0.25 Q de activación, 30 ms de tiempo de activación) sobre los 6 iones más intensos de un barrido de 400 a 1600 amu (3 µscans, 200 ms IT, 30 000 AGC), empleando exclusión dinámica. El espectro ZoomScan se empleó para la cuantificación y el espectro MS/MS para la identificación de los péptidos. El Orbitrap realizó un ciclo consistente en un barrido de 390 a 1200 amu (en trampa 1 µscan, 100 ms IT, 30000 AGC y en FTMS 1 µscan, 500 ms IT, 10^6 AGC), seguido de la fragmentación en modo CID (3 µscan, 100 ms IT, 10000 AGC, ventana de aislamiento 2 amu, 35% CE, 0.25 Q de activación, 30 ms de tiempo de activación) de los 6 iones más intensos empleando exclusión dinámica.

Para el marcaje diferencial con iTRAQ los péptidos trípticos se marcaron con el reactivo iTRAQ correspondiente (Applied Biosystems) en carbonato de trietilamonio (TEAB) 180 mM, pH 8.53. Para la muestra A se usó el marcador 116 y para la B el 117. Al cabo de 1 hora la reacción se detuvo mediante la dilución a la mitad con ácido fórmico al 0.1 % y una incubación de 30 min. Posteriormente, los péptidos se mezclaron, se limpiaron con cartuchos de intercambio catiónico SCX Oasis (Waters) empleando como solución de elución formiato amónico 1 M, pH 3, conteniendo 25 % ACN, y se desalaron en cartuchos de C18 OASIS, eluyendo los péptidos con 50% de % ACN conteniendo 0.1 % TFA. Seguidamente, los péptidos se separaron en 24 fracciones por isoelectroenfoque según se describe en el párrafo anterior, y se analizaron por RP-HPLC acoplado on-line a un LTQ u off-line a un 4800 Plus MALDI-TOF/TOF (Applied Biosystems). El LTQ se programó para fragmentar en modo PQD (4 µscans, 100 ms IT, 10000 AGC, ventana de aislamiento de 2 amu, 28 % CE, 0.6 Q de activación, 0.3 ms

tiempo de activación) los 6 iones más intensos de un barrido de 400 a 1600 amu (3 µscans, 200 ms IT, 30000 AGC), empleando exclusión dinámica. El MALDI-TOF/TOF se programó para realizar 800 disparos por espectro en un rango de masas de 850 a 3000 Da, seguido de una fragmentación en modo CID con aire (2400 disparos por espectro). En la selección de precursores se excluyeron los picos correspondientes al Glu-fibrinógeno y a la matriz (ácido α-ciano-4-hidroxicinámico), y se empleó exclusión dinámica.

Los péptidos provenientes de muestras marcadas para SILAC también se separaron por isoelectroenfoque combinado con RP-HPLC y se analizaron en LTQ-Orbitrap. En el LTQ las condiciones de scan fueron similares a las empleadas para ^{18}O, excepto que la ventana para el Zoom-scan fue de ± 8 amu.

Métodos matemáticos y software

Captura de datos de espectrometría de masas

La captura de datos crudos en los equipos Thermo (LTQ y LTQ-Orbitrap) fue realizada con el software propietario de Thermo, Xcalibur versiones 2.0, 2.0.5 y 2.0.7. Esta última versión fue utilizada para capturar datos de proteomas de iTRAQ utilizando la tecnología PQD. Para la lectura de datos del software desarrollado en este trabajo (pRatio y QuiXoT) se utilizaron las librerías de Xcalibur (Thermo) y BioWorks 3.2. Los datos crudos provenientes del equipo MALDI-TOF-TOF fueron capturados con el software propio del equipo (GPS Explorer), y los espectros de fragmentación fueron identificados con Mascot v.2.2.

Software y métodos matemáticos utilizados en análisis de datos en experimentos de identificación masiva

Motores de búsqueda y bases de datos

La gran mayoría de los espectros de fragmentación utilizados fueron analizados con el motor de búsqueda SEQUEST v.27. En el caso de los trabajos de mejora de la estimación de la False Discovery Rate se utilizó también Mascot, v.2.2, y los espectros MS^2 obtenidos con MALDI-TOF-TOF se analizaron mediante ProteinPilot.

Los parámetros de búsqueda se configuraron en función del espectrómetro de masas utilizado (la tolerancia de la masa del ión precursor depende del espectrómetro) y del tratamiento realizado a la muestra. En la Tabla MM. 1 se detallan los parámetros de búsqueda más relevantes en función del tipo de análisis realizado.

Aproximación	Equipo	tolerancia masa ión parental (Da)	tolerancia masa iones fragmento (Da)	Modificaciones opcionales (Da)	Modificaciones obligatorias (Da)
identificativa	LCQ DECA LTQ Orbitrap	2,0	1.2 1,2 1,2	Met +15.9949	Cys +57.0513
cuantitativa - 18O	LTQ Orbitrap	2,0	1,2	Met +15.9949, Lys y Arg +4.0085	Cys +57.0513
cuantitativa - SILAC	LTQ Orbitrap	2,0	1,2 1,2	Met +15.9949, Lys y Arg +6.0201	N-terminal y Lys +144.1020, Cys +57.0513
cuantitativa - iTRAQ	LTQ MALDI-TOF-TOF	2,0 0.4	1,2 0.4	Met +15.9949	

Tabla MM. 1 Condiciones de búsqueda utilizadas en los experimentos analizados.

Las búsquedas se llevaron a cabo contra la base de datos SwissProt más actualizada en cada momento del estudio. Para muestras humanas (células Jurkat y HUVEC) se empleó exclusivamente las secuencias provenientes de proteínas humanas, mientras que para muestras de levadura se utilizó una base de datos conjunta de humano y levadura, para ampliar el poder estadístico de la búsqueda.

Validación estadística de resultados de motores de búsqueda

A lo largo de este trabajo, el término "validación estadística de resultados" de identificación es referido a la generación de una lista de espectros identificados con una secuencia peptídico a la que se asocia el error cometido, expresado en términos de tasa de error (*False Discovery Rate*, FDR). La tasa de error se refiere al porcentaje de falsos que estadísticamente se encuentra en una lista de espectros identificados. Por ejemplo: de una lista de 1000 péptidos identificados con una FDR de 1% se estima que 10 de esos péptidos identificados pueden estar mal identificados. Es habitual ofrecer siempre una lista ordenada por el parámetro de calidad que se utilizó para discernir la bondad de la identificación. Dentro de este orden, podemos asociar a cada uno de los espectros identificados un valor de FDR, que se refiere entonces – y ya que la lista está ordenada – al FDR máximo que tendría la lista de espectros identificados, si el último espectro de la lista fuera aquel al que nos referimos.

SEQUEST y el método de la razón de probabilidades

Las búsquedas realizadas con el motor de búsqueda SEQUEST fueron validadas utilizando un parámetro diferente a los ofrecidos por el propio SEQUEST (Xcorr, ΔC_n , Sp…), llamado razón de probabilidades (pR) y desarrollado en nuestro laboratorio, en parte en esta tesis doctoral. El FDR fue calculado a partir de este parámetro, teniendo en cuenta para ello las estimaciones desarrolladas en esta tesis doctoral.

El método pR es un procedimiento para estimar la significatividad estadística de las identificaciones desarrollado en su inicio por Fernando Martín-Maroto y Salvador Martínez de Bartolomé. Posteriormente fue mejorado introduciendo ciertas modificaciones desarrolladas en este trabajo. La idea fundamental en la que se basa este método es utilizar correctamente las distribuciones promedio *(average score distributions)* generadas al realizar búsquedas con SEQUEST. Estas distribuciones se generan obteniendo la j-ésima mejor puntuación Xcorr de SEQUEST de todos y cada uno de los espectros contra una base de datos señuelo (*decoy*), ordenando la lista completa de puntuaciones j-ésimas, y normalizando la posición respectiva dividiéndola entre el número total de espectros analizados E. Si $\mathbf{I_N}(x)$ es la distribución generada con las primeras mejores puntuaciones y $\mathbf{H_N}(x)$ la generada con las segundas mejores

puntuaciones, entonces puede demostrarse que la probabilidad p_R de que el espectro asignado con una cierta secuencia con una primera mejor puntuación de Xcorr X_F y una segunda mejor puntuación de Xcorr X_S esté mal asignado viene dada por:

$$p_R = \frac{I_N(x_F)}{H_N(x_S)}$$

Ecuación MM. 1

Esta estimación de la probabilidad se llama razón de probabilidades. En la Figura MM. 1 se representa de forma esquemática cómo se calcula esta probabilidad. Además la distribución $\mathbf{H_N}(x)$ puede aproximarse como $\mathbf{H_N}(x) \approx \mathbf{I_N}(x)$ (véase Figura MM. 2) cuando el número de candidatos es suficientemente alto, con lo que se puede calcular la razón de probabilidades como:

$$p_R = \frac{I_N(x_F)}{I_N(x_S)}$$

Ecuación MM. 2

Para poder utilizar de forma rápida este factor, se desarrolló un software específico, llamado pRatio, que lee archivos de resultados de BioWorks 3.2 y BioWorks 3.3 (.SRF y .XML), calcula las pR de cada espectro buscado con SEQUEST y crea un archivo de resultados en el formato QuiXML (este formato se encuentra explicado en el anexo A, en la documentación explicativa de los esquemas QuiXML), de forma que este programa se integrara dentro de la plataforma de software de proteómica cuantitativa desarrollada en este trabajo.

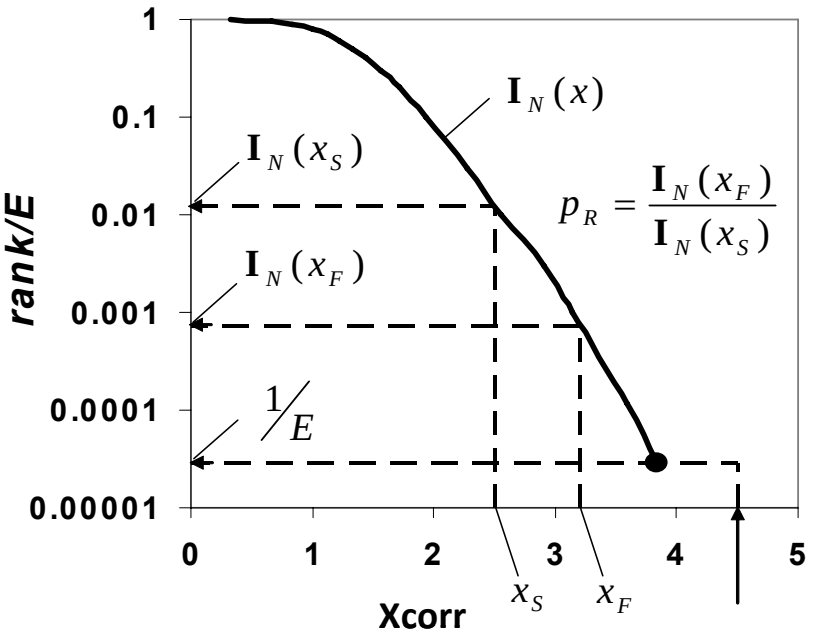


Figura MM. 1 Determinación de la razón de probabilidades. La distribución media de probabilidades de la mejor puntuación, $I_N(x)$, se obtiene buscando una colección suficientemente grande de espectros MS^2 contra una base de datos señuelo, y calculando para cada espectro su posición normalizada en el ranking de todos los espectros, según la mejor puntuación obtenida por cada uno de ellos. Los espectros se buscan después contra la base de datos objetivo, obteniéndose una mejor puntuación (xF) y una segunda mejor puntuación (xS) para cada uno de los espectros. Estas puntuaciones se interpolan numéricamente en la curva IN(x), obteniéndose las probabilidades de la mejor puntuación ($I_N(x_F)$) y de la segunda mejor puntuación ($I_N(x_S)$). La razón de probabilidades es el cociente entre estos dos valores. Cuando la mejor puntuación en la base de datos objetivo es superior a la mejor puntuación obtenida en la base de datos señuelo, únicamente se asume que la probabilidad es menor que 1/E, donde E es el número total de espectros; este procedimiento evita errores de estimación de probabilidades mediante extrapolación en una región cuya distribución es desconocida.

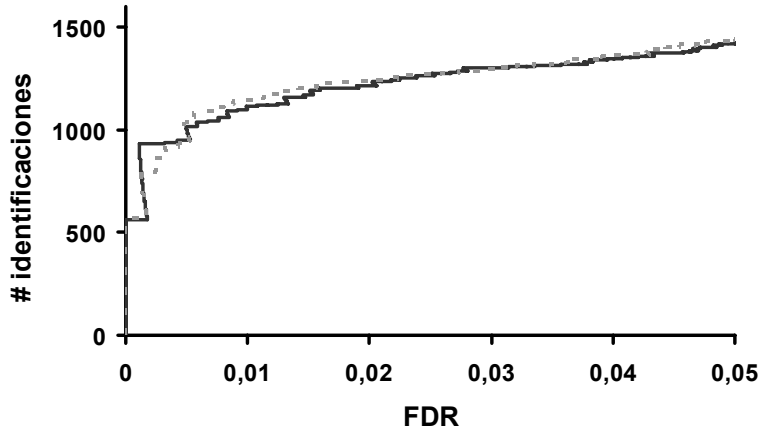


Figura MM. 2 Efecto del método utilizado para calcular la razón de probabilidades en el rendimiento de identificación de péptidos. Una extensa colección de espectros MS2 (más de 40,000) provenientes del análisis del proteoma de células Jurkat se sometió a un proceso de búsqueda contra bases de datos objetivo y señuelo. Los resultados de las búsquedas fueron analizados mediante el método de la razón de probabilidades, y se evaluó el rendimiento de la identificación de péptidos representando el número de péptidos identificados en función de la tasa de error FDR. La razón de probabilidades se calculó como $I_N(x_F)/I_N(x_S)$ (línea negra), o como $I_N(x_F)/H_N(x_S)$ (línea gris).

Mascot

Las búsquedas realizadas con el motor de búsqueda Mascot fueron validadas utilizando como parámetro el generado por el motor de búsqueda *probability based Mowse Score*, y la estimación de la FDR fue llevada a cabo según se expone en los resultados de este trabajo.

Software y métodos matemáticos utilizados en análisis de datos de proteómica cuantitativa mediante marcaje con isótopos estables

Métodos de programación

Todo el software desarrollado fue programado en el ámbito de Microsoft .NET, específicamente en C#. El *framework* de .NET utilizado fue en inicio la versión 2.0, pero se fue actualizando la versión utilizada, siendo la versión final la 3.5, y todos los proyectos anteriores a esta versión fueron reconvertidos a ésta.

En el diseño de software, se intentó en la medida de lo posible utilizar patrones de diseño de *Gang of Four* (E Gamma, 1995) (GoF).

Algunas versiones de prueba fueron inicialmente programadas, por comodidad, como macros de Visual Basic en Microsoft Excel, y, una vez validadas, fueron posteriormente traducidas e incorporadas a los proyectos .NET correspondientes.

Software de cuantificación QuiXoT

Para los análisis de proteómica cuantitativa se desarrolló y utilizó como banco de pruebas una plataforma de software específica para proteómica cuantitativa, cuyo eje central es el programa QuiXoT, un programa que adquiere datos de cualquier tipo de espectrómetro de masas y cuantifica en varios tipos de marcaje isotópico pre-establecidos (^{18}O, SILAC y iTRAQ), permitiendo también configurarlo para realizar cuantificaciones sobre cualquier tipo de marcaje isotópico. El desarrollo de este software y de toda la plataforma está descrito en el capítulo de resultados "Integración de una plataforma de software de proteómica cuantitativa (QuiXoT)".

Características concretas del software de cuantificación no desarrolladas en esta tesis doctoral

Algoritmo de estimación de expresión diferencial utilizando marcaje ^{18}O

El algoritmo que se utiliza para la cuantificación de péptidos marcados con ^{18}O se había desarrollado previamente en nuestro laboratorio (A Ramos-Fernandez et al., 2007). Este algoritmo se basa en el cálculo de la concentración de péptidos de las especies A (especie de péptidos no marcada), B_2 (especie marcada con dos ^{18}O en su extremo carboxilo), B_1 (especie marcada con sólo un ^{18}O) y B_0 (especie que corresponde a la muestra marcada, pero que no tiene incorporado ningún ^{18}O en su extremo carboxilo) teniendo en cuenta las reacciones de incorporación y pérdida de ^{18}O en el extremo carboxilo del péptido:

$$B_0 \underset{p\alpha k}{\overset{k}{\rightleftarrows}} B_1 \underset{pk}{\overset{\alpha k}{\rightleftarrows}} B_2$$

Ecuación MM. 3

Mediante un análisis cinético de esta reacción, puede demostrarse que la concentración de péptido total de la muestra marcada B = B_0+B_1+B_2 puede descomponerse en función de la concentración de cada una de las especies de la muestra marcada, utilizando un parámetro f que evalúa la eficiencia de marcaje (A Ramos-Fernandez et al., 2007) :

$$B_2 = Bf^2$$
$$B_1 = B \cdot 2f(1-f)$$
$$B_0 = B(1-f)^2$$

Ecuación MM. 4

Por lo que mediante un algoritmo de resolución de problemas no lineales por mínimos cuadrados, como Newton-Gauss, pueden encontrarse soluciones simultáneas para las concentraciones de péptido A y B y la eficiencia de marcaje f, que expliquen correctamente la distribución isotópica de los espectros correspondientes a péptidos marcados.

Corrección isotópica en el marcaje de iTRAQ

Los iones reporteros de iTRAQ 4-plex están separados entre sí por 1 Da. En función de la pureza isotópica de los reactivos respectivos se produce una distribución isotópica para cada una de las especies de los iones reporteros. Estas distribuciones solapan con la de los iones reporteros vecinos, por lo que se hace necesaria una corrección isotópica que tenga en cuenta la intensidad detectada de cada uno de los iones reportero en función de la pureza de cada uno de los compuestos con los que se realiza el marcaje (facilitada por el fabricante). Se obtiene un sistema de ecuaciones lineales que se resuelve aplicando la regla de Cramer, cuyo determinante se calcula utilizando el método propuesto por Shadforth(IP Shadforth et al., 2005). Todos estos cálculos se implementaron en el software de cuantificación desarrollado en esta tesis.

Detección de máximos en picos cromatográficos mediante el software QuiXtoQuiX

En los experimentos analizados mediante Orbitrap se realizó una detección de máximos de los picos cromatográficos de los péptidos identificados mediante espectros MS/MS, de forma que la cuantificación se realizara sobre los espectros más intensos de cada pico cromatográfico. Para ello se utilizó un software propio del laboratorio, QuiXtoQuiX, programado por Marco Trevisan. Este software procesa los archivos de datos adquiridos .RAW de Thermo, buscando el máximo del pico cromatográfico al que pertenece cada identificación reseñada en un archivo QuiXML, añadiendo esta información a este mismo archivo QuiXML. Esta herramienta se encuentra integrada dentro de la plataforma de software QuiXoT.

Resultados

1. Desarrollo de nuevos métodos para el análisis estadístico de resultados en experimentos de identificación masiva de proteínas

1.1 Optimización del método de la Razón de Probabilidades para el análisis de resultados obtenidos con el motor de búsqueda SEQUEST

Corrección de la Razón de Probabilidades en función de la carga y de la masa del ión parental.

Como se describe en la introducción, en un trabajo previo de Fernando Martín Maroto se realizó un análisis matemático de las distribuciones promedio de puntuaciones de SEQUEST. De este análisis se obtuvo la predicción teórica de que la razón de probabilidades (pR) es el indicador óptimo para tener en cuenta conjuntamente la información contenida en las dos primeras puntuaciones del motor de búsqueda SEQUEST (lo que es equivalente a la información contenida en los parámetros Xcorr y ΔC_n, puesto que la segunda mejor puntuación puede calcularse a partir de ellos). El algoritmo fue probado de forma preliminar y sin ningún tipo de optimización por Salvador Martínez de Bartolomé, con resultados muy prometedores, ofreciendo un rendimiento en experimentos de identificación masiva de péptidos (en términos del número de asignaciones péptido-espectro, o PSMs, validadas para la misma FDR) comparable, aunque no claramente superior a otros métodos ya publicados (JE Elias,SP Gygi, 2007, A Keller et al., 2002).

Aunque el modelo matemático desarrollado por Fernando Martín-Maroto predice que el indicador pR tiende a la unidad cuando las PSM son aleatorias (es decir cuando se identifica un péptido falso) y esta propiedad es independiente de cualquier otra característica que intervenga en la asignación de las PSM, el modelo no da información sobre la distribución estadística del parámetro pR en torno a ese valor. Nos planteamos la posibilidad de que al tratarse de un cociente de valores, la forma de la distribución pudiera depender de la magnitud de dichos valores, mostrando una dispersión mayor cuanto menores fueran los valores usados para calcularla. Esta dispersión podría producir un efecto sobre la estadística de orden, al coexistir datos con dispersiones diferentes, y podría influir en la asignación de identificaciones correctas, ya que éstas se asocian a valores de pR alejados del punto central, precisamente en la “cola” de la distribución.

Para analizar este efecto, realizamos un estudio detallado del comportamiento de pR en función de los principales factores que afectan a la magnitud numérica de la puntuación SEQUEST (Xcorr), la carga del ión parental y la masa del péptido (A Keller et al., 2002). Para ello

analizamos por separado las distribuciones de los iones con carga z = 2 y z = 3 y, por otro lado, las distribuciones correspondientes a tres rangos de masa del ión parental (de 800 a 1300 Da, de 1300 a 1800 Da y de 1800 a 2300 Da). Como se observa en la (Figura R. 1), los valores de pR se distribuyeron en todos los casos en torno a la unidad, de acuerdo al modelo teórico. Sin embargo, en dichas figuras detectamos ligeras diferencias en los extremos de las distribuciones de pR separadas por masa (Figura R. 1B) y diferencias muy claras en aquéllas separadas por carga (Figura R. 1A). Estas diferencias producen un comportamiento general subóptimo del parámetro pR a la hora de asignar las PSM correctas, ya que el parámetro pR actúa como un clasificador más sensible en unos casos que en otros.

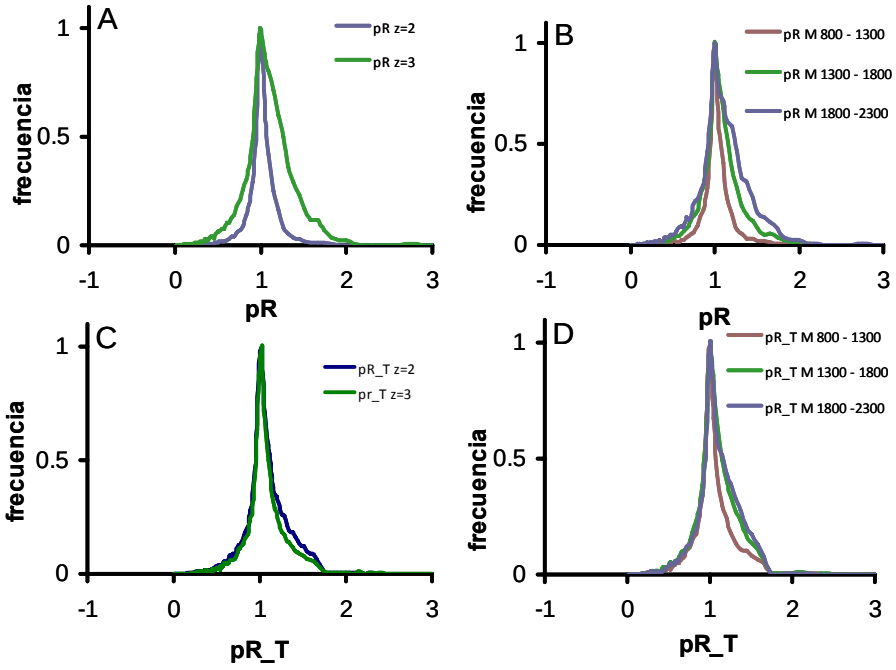


Figura R. 1 Análisis de las distribuciones de la razón de probabilidades en función de la carga y la masa del ión parental. Se representa la distribución (normalizada a la unidad) de los valores de la razón de probabilidades obtenidos en un experimento que generó 40,000 espectros MS^2 (proteoma de células Jurkat) clasificando los datos de acuerdo a la carga del ión parental (paneles A y C) o a la masa del péptido (paneles B y D). En A y B se estimaron las razones de probabilidades utilizando la puntuación Xcorr calculada por SEQUEST. En C y D se calcularon las distribuciones usando las puntuaciones Xcorr de SEQUEST previamente corregidas para tener en cuenta los factores carga y masa, según se explica en el texto.

Para corregir este efecto se introdujo un factor de corrección en función de carga (z) y masa (M) del ión parental. La corrección se aplicó transformando el valor de la puntuación ofrecida por SEQUEST (Xcorr) de la siguiente manera:

$$Xcorr_T\left(Xcorr, M, R\right) = \frac{\log_{10}\left(Xcorr/R\right)}{\log_{10}\left(2 \cdot M/M_0\right)} \quad ,$$

$$donde \qquad R = \begin{cases} 1, & si \quad z = 2 \\ 1.22, & si \quad z = 3 \end{cases},$$

Ecuación R. 1

donde $Xcorr_T$ es la puntuación SEQUEST corregida y M_0 es la masa promedio de los aminoácidos (M_0 = 110 Da). En esta ecuación la corrección en masa es una normalización logarítmica de la longitud del péptido estimada mediante M/M_0 y está basada en la corrección empírica sugerida por Nesvizhskii y colaboradores(AI Nesvizhskii et al., 2006).

En cuanto a la corrección de carga, se tuvo en cuenta el rango de masa/carga en el que la señal de ruido de fondo del espectro MS/MS pudiera ser erróneamente asignada por el motor de búsqueda a alguno de los fragmentos teóricos del péptido. Este razonamiento se ilustra en la Figura R. 2. Los iones monocargados sólo pueden generar fragmentos con carga 1+, que se distribuyen a lo largo de la zona marcada en azul (zona Δy^{+}). Los iones bicargados generan fragmentos tanto de carga 1+, que abarcan la zona azul Δy^{+}, como de carga 2+, que abarcan la zona naranja Δy^{2+}. Análogamente, para los iones tricargados hay que tener en cuenta la zona azul Δy^{+} (fragmentos monocargados), la zona naranja Δy^{2+} (fragmentos bicargados) y la zona roja Δy^{3+} (fragmentos tricargados). Teniendo en cuenta que

$$\Delta y^{2+} = \Delta y^{+} / 2 \quad , \quad \Delta y^{3+} = \Delta y^{+} / 3$$

Ecuación R. 2

el tamaño del espacio de probabilidades en el que una señal cualquiera del espectro MS/MS puede ser asignada al azar a ión teórico de una secuencia peptídica dada es una función de la carga del ión parental de acuerdo a la siguiente ecuación:

$$Esp(z=1) = \Delta y^{+}$$
$$Esp(z=2) = \Delta y^{+} + \Delta y^{2+} = \Delta y^{+} \cdot (1 + 1/2)$$
$$Esp(z=3) = \Delta y^{+} + \Delta y^{2+} + \Delta y^{3+} = \Delta y^{+} \cdot (1 + 1/2 + 1/3)$$

Ecuación R. 3

En esta transformación sólo hemos tenido en cuenta iones de cargas 2 y 3, ya que son las especies de péptidos digeridos con tripsina más abundantes que se generan por ionización ESI. Finalmente, la proporción entre los espacios de probabilidades de estas dos cargas toma el siguiente valor

$$\frac{Esp(z=3)}{Esp(z=2)} = \frac{\Delta y^{+} \cdot (1 + 1/2 + 1/3)}{\Delta y^{+} \cdot (1 + 1/2)} = \frac{11}{9} \approx 1.22$$

Ecuación R. 4

que justifica la corrección introducida en la Ecuación R. 1.

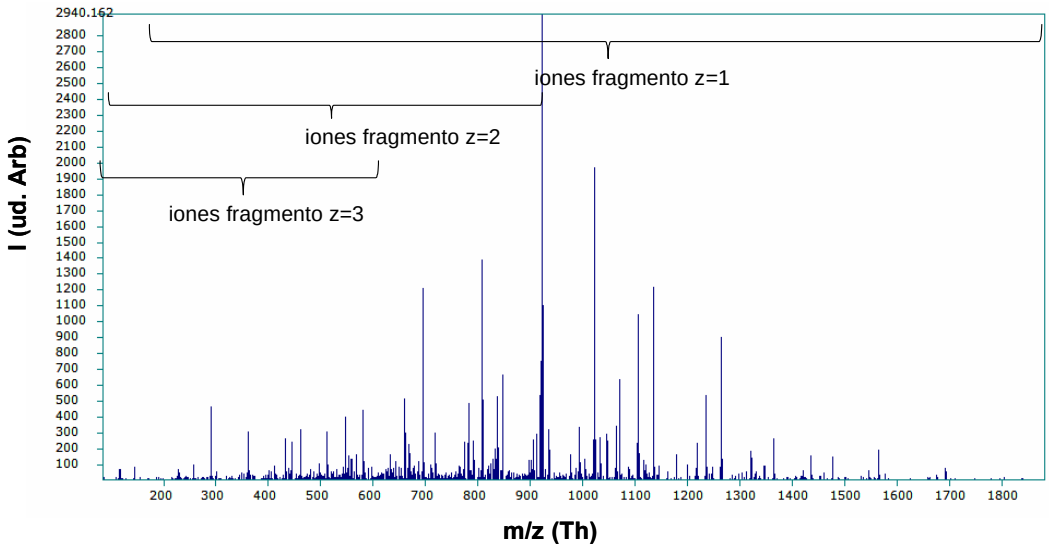


Figura R. 2 Esquema de la corrección de puntuaciones Xcorr en función de la carga del ión parental. En el espectro aparecen señaladas las regiones en las que pueden aparecer los iones correspondientes a las series *y* y *b cuando la carga del péptido es* z=1, z=2 ó z=3.

En la Figura R. 3 se representan las distribuciones normalizadas de la puntuación de SEQUEST sin transformar (Xcorr) y la transformada ($Xcorr_T$). En esta figura se observa un claro desplazamiento de las distribuciones Xcorr en función tanto de la carga (Figura R. 3 A) como de la masa (Figura R. 3 B), efecto ya notado por muchos autores en trabajos previos. En dicha figura también observamos que las distribuciones de la puntuación transformada están mucho más centradas (Figura R. 3 C y D). Sin embargo los extremos correspondientes a puntuaciones más altas (donde se sitúan los valores correspondientes a las asignaciones correctas) siguieron mostrando diferencias claras, ilustrando la dificultad de normalizar un parámetro de desarrollo empírico como el usado por SEQUEST. Este resultado explica en parte por qué los principales métodos de inferencia estadística basados en este indicador utilizan distribuciones de Xcorr separadas en función de la carga (A Keller et al., 2002) En claro contraste, las distribuciones de pR obtenidas usando las puntuaciones transformadas son prácticamente indistinguibles tanto en función de la carga (Figura R. 1 C) como de la masa (Figura R. 1 D), mostrando cómo el indicador pR, una vez corregido el efecto numérico (aunque sea de forma incompleta) debido a la puntuación de SEQUEST, tiende a hacerse independiente de otros factores que influyen en las asignaciones, de acuerdo a las predicciones del modelo teórico.

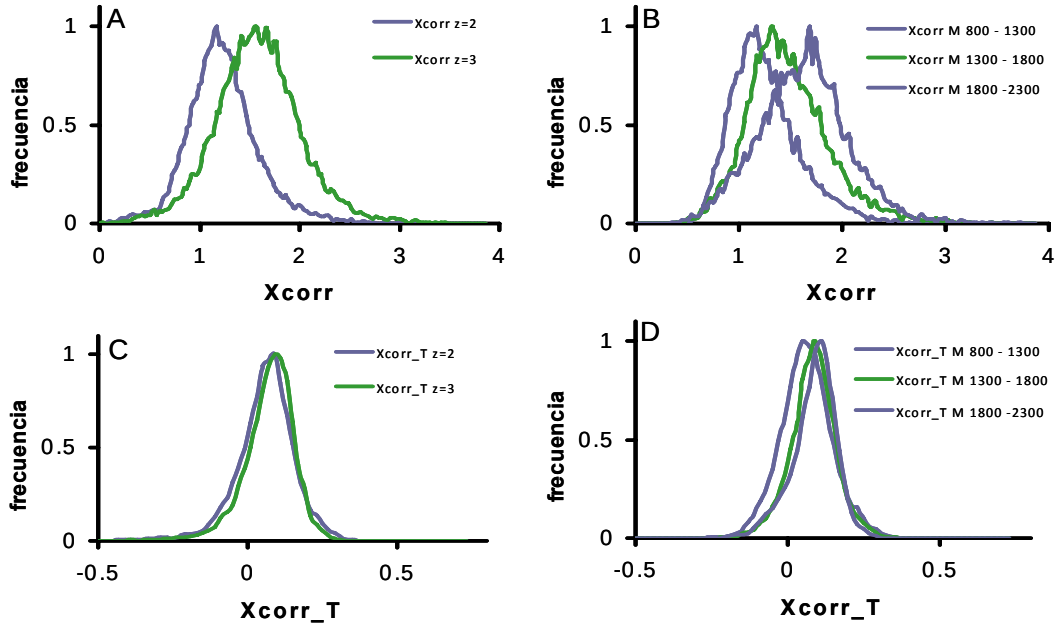


Figura R. 3 Análisis de las distribuciones medias de la mejor puntuación Sequest (Xcorr) en función de la carga y de la masa del ión parental**. Se utilizó un experimento de 40,000 espectros MS^2 (proteoma de células Jurkat) separando las puntuaciones Xcorr de acuerdo a la carga del ión parental (paneles A y C) y a la masa (paneles B y D). En A y B se utilizaron las puntuaciones Xcorr ofrecidas por SEQUEST. En C y D se utilizaron las puntuaciones Xcorr de SEQUEST corregidas para tener en cuenta los factores carga y masa, respectivamente.**

En la Figura R. 4 se comparó el número de PSMs obtenidas usando las puntuaciones sin corregir o corregidas por el efecto de la masa y de la carga, en función de la FDR estimada. En dicha figura se observa cómo la pR calculada usando puntuaciones corregidas tiene un rendimiento claramente superior. El Xcorr corregido mejoró también a su homólogo sin corregir, y a valores de FDR bajas (menores al 2.5%), la pR obtenida de datos sin corregir consigue un número de identificaciones aún superior que Xcorr corregido. En la Figura R. 5 se comparan las prestaciones del indicador pR (usando Xcorr corregidas o no) con otros métodos descritos en la literatura, observándose que la pR usando puntuaciones corregidas tiene unas prestaciones superiores al método empírico 2VG desarrollado previamente en nuestro laboratorio (D Lopez-Ferrer et al., 2004) y también al método basado en la optimización iterativa de XCorr y ΔC_n. Nótese que el método 2VG ha sido previamente comparado con otros métodos descritos en la literatura, usando los mismos datos que en este trabajo, y se ha demostrado tener unas prestaciones ligeramente superiores a PeptideProphet (D Lopez-Ferrer et al., 2004).

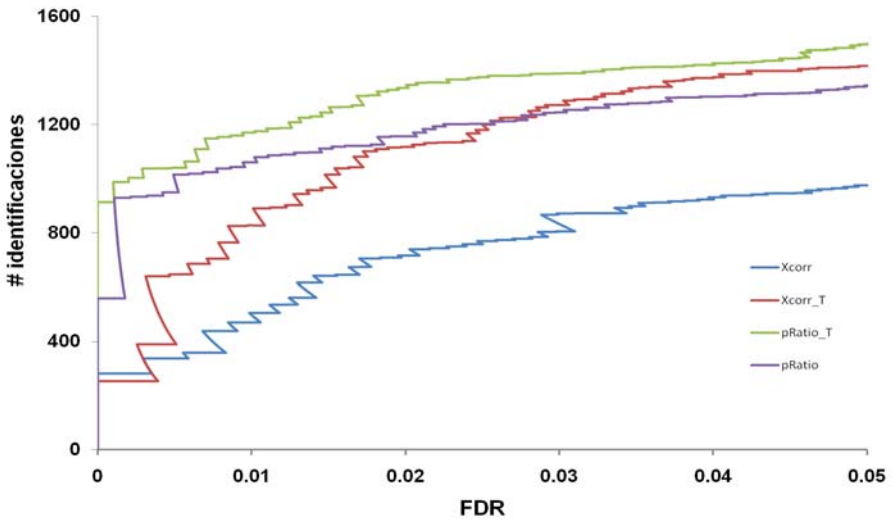


Figura R. 4 Comparativa del rendimiento de identificaciones obtenido usando diferentes indicadores. Se representó el número de péptidos identificados frente a la tasa de error de los resultados obtenidos usando como criterio cada uno de los indicadores siguientes: (azul) Xcorr; (rojo) Xcorr corregido para tener en cuenta los factores carga y masa; (violeta) pR, y (verde) pR corregido para tener en cuenta los factores carga y masa.

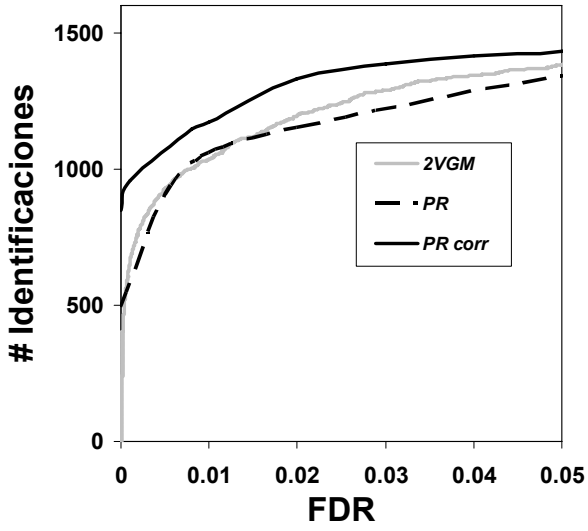


Figura R. 5 Comparativa del rendimiento de identificaciones obtenido usando diferentes indicadores. Se representó el número de péptidos identificados frente a la tasa de error de los resultados obtenidos usando como criterio cada uno de los indicadores siguientes: (gris oscuro) método de la distribución gaussiana en dos dimensiones (2VGM), (gris claro) método iterativo de optimización de umbrales de Xcorr y ΔC_n, (negro, línea discontínua) pR calculada usando puntuaciones Xcorr no corregidas por el efecto de la masa y de la carga, y (negro, línea contínua) pR usando puntuaciones Xcorr corregidas.

Integración del punto isoeléctrico.

En estudios de identificación masiva de proteínas, uno de los métodos más eficaces para fraccionar los péptidos generados por la digestión de un proteoma, de forma previa a su análisis por RP-HPLC en tándem con espectrometría de masas, es la separación mediante isoelectroenfoque en fase líquida usando geles lineales con gradiente de pH inmovilizado. Este método permite además determinar el punto isoeléctrico de los péptidos (pI), que es un parámetro característico de su secuencia y es independiente de la información obtenida por espectrometría de masas. La información del pI puede aprovecharse para mejorar la validación de las identificaciones realizadas por motores de búsqueda.

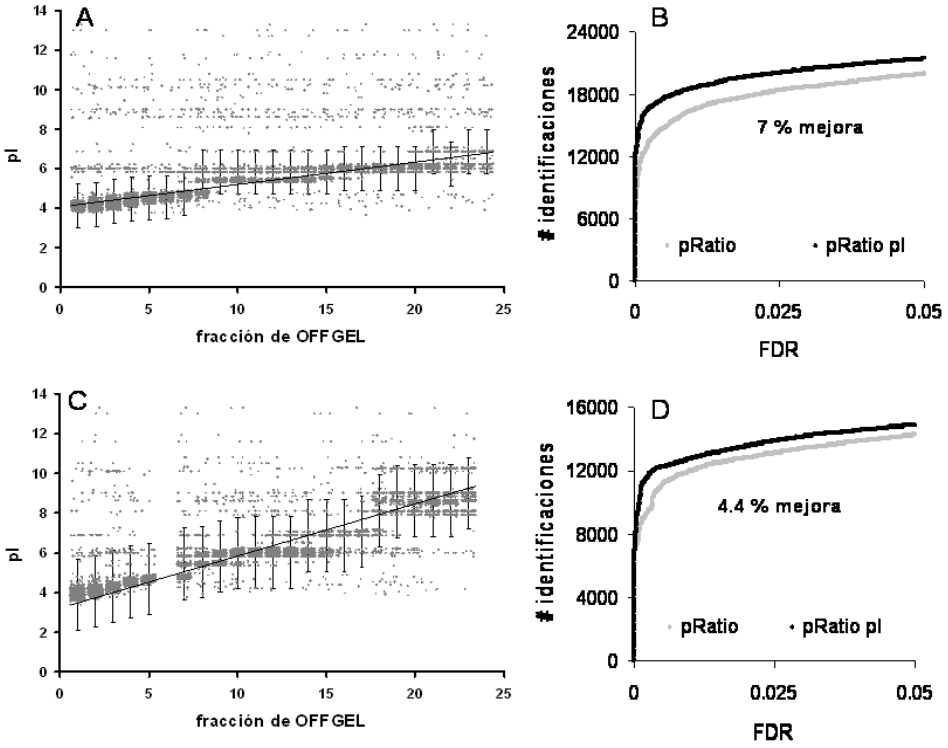


Figura R. 6 Utilización del punto isoeléctrico para mejorar el rendimiento de identificación de péptidos a partir de espectros MS^2. Los paneles A y C representan el punto isoeléctrico de los péptidos frente a la fracción Offgel en la que se identifican usando tiras de pH de 4 a 7 (A) o de 3 a 10 (C). Las identificaciones se asignaron con una tasa de error FDR<5% utilizando la razón de probabilidades; las barras de error indican el ancho de pI estimado de forma que maximiza el número de identificaciones conseguidas en función de la FDR. Los paneles B y C muestran las correspondientes gráficas de rendimiento de identificación obtenidas utilizando exclusivamente la razón de probabilidades (en gris), o utilizando la razón de probabilidades combinada con el criterio de punto isoeléctrico (en negro).

Para poner a punto un algoritmo que permita mejorar la identificación de péptidos con la ayuda de la información del pI, se utilizaron dos proteomas de prueba diferentes. Ambos proteomas fueron digeridos, y los péptidos resultantes fueron separados mediante isoelectroenfoque en un sistema de OffGel, el primero de ellos en un rango de pH de 4 a 7 y el segundo en un rango de pH de 3 a 10. En la Figura R. 6 A y C se representan los puntos isoeléctricos calculados para las secuencias peptídicas identificadas usando el método de la pR con una tasa de error de FDR<0.01, en función de la fracción de OffGel en las que se identificó cada péptido. Se observa claramente que los péptidos se separan a lo largo de las fracciones de acuerdo a su pI, lo que sugiere que aquéllos cuyo pI no corresponda al resto de los péptidos de la misma fracción podrían tratarse de identificaciones falsas.

Para combinar la razón de probabilidades (pR) y el factor de punto isoeléctrico (pI), debe tenerse en cuenta que la pR es la probabilidad de que la asignación péptido – espectro (o PSM) se deba al azar, es decir, que sea incorrecta, y a este factor habría que añadir algún parámetro que evaluara si el pI del péptido fuera o no el esperado según la fracción IEF en la que se identificara. Por tanto, y puesto que se trata de sucesos independientes, la probabilidad combinada de que una asignación sea correcta teniendo en cuenta conjuntamente la información generada por espectrometría de masas y el punto isoeléctrico, P_{pR_pI} podría expresarse de la siguiente manera

$$P_{pR\_pI} = (1 - pR) \cdot F_{pI}$$

Ecuación R. 5

siendo F_{pI} una función dependiente del pI del péptido identificado y de la fracción en la que se encuentra, que evalúa la probabilidad de que el pI del péptido identificado sea correcto. La elección de la F_{pI} más adecuada se realizó buscando el algoritmo óptimo que maximizara el rendimiento de identificaciones de péptidos, es decir, el número de identificaciones observadas para la misma FDR.

En una primera aproximación asumimos que la distribución de pI de los péptidos en cada una de las fracciones sigue una función gaussiana. Para determinar esta función se estimó el valor medio de pI de los péptidos de cada fracción y la varianza global de los péptidos identificados en todas las fracciones en torno al valor medio de cada una de ellas. En otras palabras, asumimos que el pI de todos los péptidos tiene la misma varianza, independientemente de la fracción donde se encuentran. Con esta función se determinó la probabilidad de que el punto isoeléctrico del péptido identificado se encontrase en la

correspondiente fracción. Sin embargo, cuando se utilizó la probabilidad conjunta dada por la Ecuación R. 5 no detectamos mejoras apreciables en el rendimiento de identificación en comparación con los resultados obtenidos usando sólo el parámetro pR (datos no mostrados). Ello se debe a que el pI de muchos de los péptidos identificados correctamente se desvía con frecuencia del pI medio de la fracción, y este efecto penaliza la identificación correspondiente, compensando la mejora que se obtiene al usar la información dada por el pI.

Por esta razón buscamos una función más sencilla que no penalizara las identificaciones correctas cuando el pI del péptido sólo se desviara ligeramente respecto al valor medio esperado. El pI en torno al valor medio de cada fracción lo modelamos usando una función cuadrada centrada en dicho valor, de ancho σ, es decir

$$F_{pI}\left(pI_{seq},\mu,\sigma\right)=\begin{cases}1, si\left|pI_{seq}-\mu\right|\le\sigma\\0, si\left|pI_{seq}-\mu\right|>\sigma\end{cases}$$

Ecuación R. 6

donde μ es el valor medio del pI de cada fracción. Con este criterio, aquellas identificaciones que no se encuentren dentro de un margen aceptado σ de punto isoeléctrico son consideradas falsas.

El algoritmo primero identifica los péptidos usando sólo la pR como criterio, con una FDR menor del 1% y calcula su pI. Los valores de pI de esta población de péptidos se usan entonces para estimar los valores medios de pI de cada fracción usando la mediana como estimador más robusto que la media. Posteriormente el ancho de la función σ se optimiza mediante un algoritmo iterativo que parte de la solución más conservadora (σ = 7, que en términos de ancho de pI puede considerarse infinito y que equivale a usar sólo la pR como criterio de identificación) y va disminuyendo su valor hasta obtener el máximo número de identificaciones a una FDR preestablecida. De esta manera la aplicación del criterio de pI nunca da un resultado peor que el obtenido usando sólo la pR.

En los proteomas analizados utilizando este nuevo factor se obtuvo un rendimiento significativamente superior de identificaciones (un 7% en el proteoma separado en el rango de pH 4 a 7, y un 4.4% en el proteoma separado en el rango de pH de 3 a 10) (Figura R. 6 B y C). Este resultado típico lo hemos confirmado en numerosos experimentos llevados a cabo en nuestro laboratorio (no se muestran). En algunos casos hemos observado que el método es muy robusto frente a posibles problemas surgidos durante la separación de punto isoeléctrico,

puesto que en el peor de los casos, el algoritmo siempre escoge la solución con la que obtenga más identificaciones. El algoritmo ha sido integrado en un programa que realiza las validaciones de las identificaciones de SEQUEST mediante el método de la pR.

1.2 Método refinado para la estimación de la tasa de error (False Discovery Rate) en experimentos de identificación de péptidos a gran escala

El criterio de validación más aceptado en identificación de péptidos en experimentos masivos es la FDR (H Choi,AI Nesvizhskii, 2008, JE Elias,SP Gygi, 2007). Para obtener el máximo rendimiento de identificaciones válidas y tener la certeza de que los resultados que manejamos son correctos, una estimación correcta de la FDR es crucial, pero los dos métodos de cálculo de la FDR más extendidos actualmente adolecen de ciertos problemas de sobreestimación de la FDR, como ya se comentó en la Introducción.

Decidimos analizar en mayor profundidad los métodos actuales de estimación de la FDR y buscar una solución correcta a los problemas de estimación de este parámetro. Para ello, analizamos el comportamiento de los motores de búsqueda representando los resultados de las mejores puntuaciones de espectros (PSMs) obtenidas en una base de datos objetivo, o *target*, en función de las PSMs obtenidas al realizar las búsquedas de los mismos espectros contra una base de datos aleatoria, señuelo, o *decoy(M Fitzgibbon et al., 2008)*. En la Figura R. 7 se representó el comportamiento de las puntuaciones de dos motores de búsqueda muy conocidos (SEQUEST y MASCOT). En el caso de SEQUEST se analizaron las puntuaciones Xcorr y el parámetro pR. Los puntos bajo la línea diagonal representan a aquellos cuya mejor puntuación en la búsqueda en la base de datos objetivo es más alta que en la búsqueda en la base de datos señuelo. Asumiendo una base de datos señuelo ideal, aquellos espectros cuya identificación no sea correcta deberían obtener una puntuación máxima equiprobable en ambas bases de datos. En otras palabras, la nube de puntuaciones aleatorias debe distribuirse de forma simétrica en torno a la diagonal principal (Figura R. 7).

Para calcular el *false discovery rate* (FDR) se estableció un cierto umbral de puntuación **t**, que delimita la población de PSM asignadas de la que se desea conocer su FDR. Al aplicar este umbral tanto a las puntuaciones en las búsquedas objetivo como en las señuelo, aparecen cuatro categorías en las que se pueden clasificar los PSMs que superan la puntuación umbral **t**: los PSMs que solamente superan la puntuación **t** en la búsqueda objetivo (*target only*, **to**), los que solamente superan la puntuación **t** en la búsqueda señuelo (*decoy only*, **do**), aquellos que superan la puntuación **t** en la búsqueda objetivo pero no en la señuelo (*target better*, **tb**), y

finalmente aquellos que superan la puntuación **t** en la búsqueda señuelo pero no en la objetivo (*decoy better*, **db**).

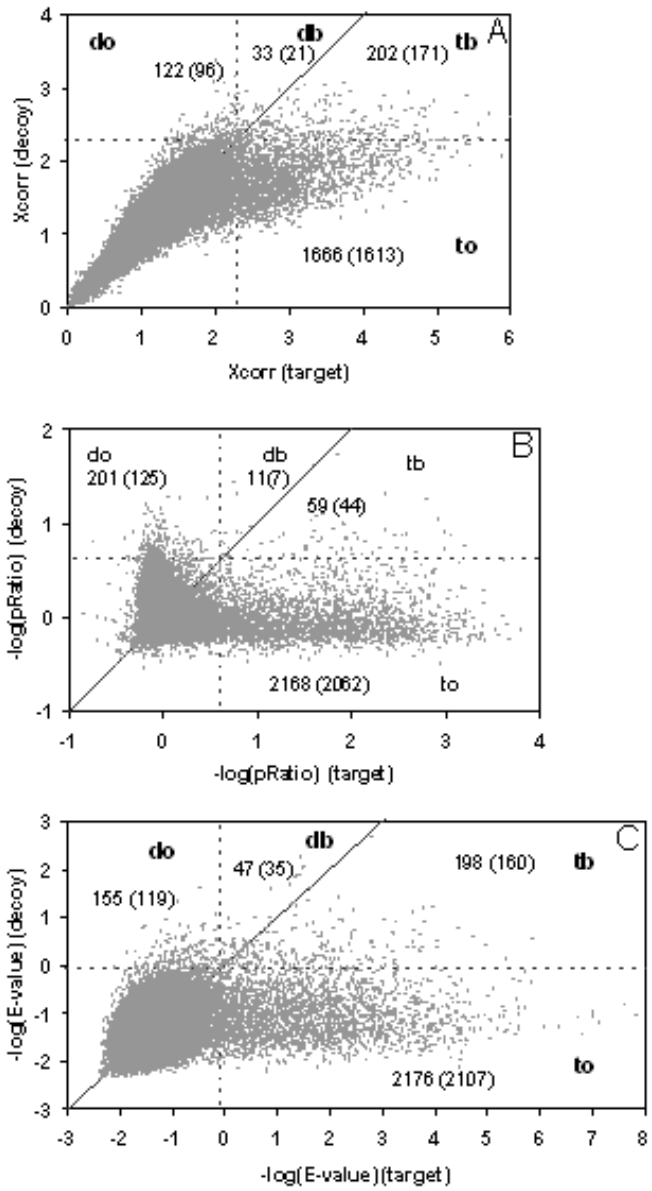


Figura R. 7 Representación de la distribución conjunta de tres tipos diferentes de puntuaciones en bases de datos objetivo y señuelo. En estas gráficas se representa la puntuación obtenida en la base de datos señuelo frente a la obtenida en la base de datos objetivo para cada uno de los péptidos. Las puntuaciones utilizadas son: (A) Xcorr de SEQUEST; (B) razón de probabilidades de puntuaciones SEQUEST; y (C) E-values de Mascot. En (B) y (C) se representan los datos en escala logarítimica. Las letras se refieren a las seis regiones de la distribución conjunta delimitadas por el umbral de puntuación (en líneas discontínuas) y la línea diagonal, tal y como se describe en el texto.

En el caso de realizar dos búsquedas separadas, una en la base de datos objetivo y otra en la base de datos señuelo, para estimar la FDR se divide el número de puntos obtenidos por encima del umbral **t** en la búsqueda señuelo, es decir, db+tb+do (la estimación de PSMs falsos) entre el número de puntos que se han encontrado por encima del umbral **t** en la búsqueda objetivo, es decir, db+tb+to. Por tanto, la estimación de la FDR utilizando bases de datos separadas, en nuestro esquema de regiones vendría dada por:

$$FDR_{SD} = \frac{db + tb + do}{db + tb + to}$$

Ecuación R. 7

La forma de la Ecuación R. 7 permite apreciar inmediatamente que este método tiene en cuenta para la estimación de PSMs falsos la población **tb**, que puede contener una población importante de PSMs correctamente identificados, con lo que se sobreestima la FDR de forma sistemática. Para evitar este efecto, y tal como proponen Elias y Gygi(JE Elias,SP Gygi, 2007), se puede realizar una única búsqueda contra una base de datos concatenada que incluya tanto la base de datos objetivo como la base de datos señuelo, de forma que se establece un sistema de competición entre los resultados de ambas bases de datos. La población que debe tomarse en este método como población de PSMs identificados es **do**+**db**+**tb**+**to**, ya que al realizarse la búsqueda conjuntamente, los PSMs pertenecientes a la base de datos objetivo y los de la señuelo no pueden ser separados. La estimación de falsos es de dos veces **do**+**db**, utilizando la simetría entre resultados objetivo-señuelo alrededor de la diagonal. Por tanto, la estimación de la FDR utilizando bases de datos concatenadas, en nuestro esquema de regiones, vendría dada por:

$$FDR_{CD} = \frac{2 \cdot (do + db)}{do + db + tb + to}$$

Ecuación R. 8

Este análisis por zonas da el fundamento para diseñar una estimación refinada de la FDR, aprovechando de forma correcta la simetría objetivo-señuelo. La población sobre la que se desea estimar la FDR debe estar formada exclusivamente por todos aquellos PSMs de la búsqueda objetivo que superan la puntuación umbral **t**, es decir, **db**+**tb**+**to**. Para estimar los falsos positivos presentes en esta población, hemos de incluir las poblaciones **do** (para estimar los falsos positivos de **to**), y dos veces **db** (para estimar los falsos positivos presentes en tanto

db como **tb**). Por tanto, la estimación de la FDR utilizando este método refinado vendría dada por:

$$FDR_{TD} = \frac{do + 2 \cdot db}{db + tb + to}$$

Ecuación R. 9

Comparando la Ecuación R. 9 con las Ecuación R. 7 y R.8 se deduce que esta estimación de FDR produce siempre un valor inferior para cualquier valor de las 4 poblaciones que se tienen en cuenta para cualquiera de las estimaciones (FDR_{SD} o FDR_{CD}), como puede observarse en la Figura R. 8, realizada para la misma colección de datos utilizada en la Figura R. 7. Por ello, para un mismo conjunto de datos, siempre se consigue una estimación de la FDR igual o mejor mediante el método refinado, lo que en términos generales aumenta el número de identificaciones (para un mismo valor de FDR). Se concluye que el método refinado es más sensible que los anteriores.

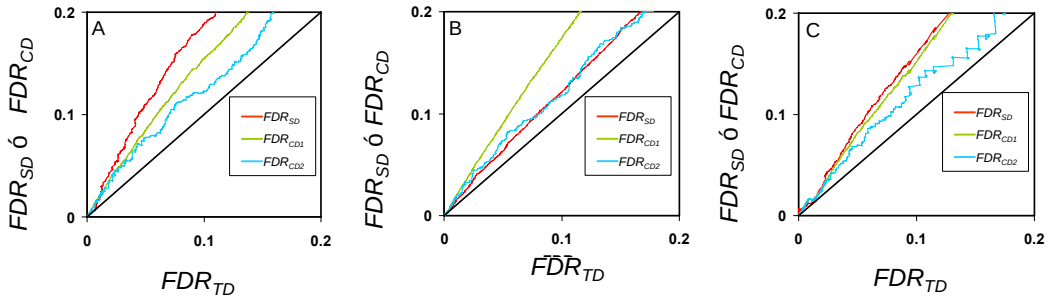


Figura R. 8 Comparación de las tasas de error obtenidas utilizando la aproximación por bases de datos separadas (FDR_{SD}, líneas rojas) y la aproximación de bases de datos concatenadas (FDR_{CD}, líneas verdes y azules) con las tasas de error obtenidas con el método refinado propuesto en este trabajo (FDR_{TD}). (A), (B) y (C) se refieren a las tres puntuaciones descritas en la figura anterior. Nótese que hay dos curvas de FDR_{CD} en cada figura; una de ellas (FDR_{CD1}) se obtiene utilizando como criterio el mismo umbral de puntuación (líneas verdes), y el otro (FDR_{CD2}) se obtiene utilizando como criterio el mismo número de identificaciones positivas (líneas azules), según se explica en el texto.

Los tres desarrollos expuestos en los apartados anteriores, ofrecen conjuntamente resultados óptimos en la inferencia estadística de identificaciones válidas de espectros MS^2 obtenidas en experimentos de alto rendimiento, consiguiendo un número de identificaciones (utilizando la misma tasa de error) claramente superior(S Martinez-Bartolome et al., 2008) a los descritos en la bibliografía, entre los que pueden destacarse Peptide Prophet(A Keller et al.,

2002) y métodos empíricos basados en el ajuste de distribuciones gaussianas como el método 2VG(D Lopez-Ferrer et al., 2004).

2. Desarrollo de un modelo estadístico universal para el análisis de datos de proteómica cuantitativa basada en el marcaje con isótopos estables.

2.1 Un modelo estadístico para la cuantificación de péptidos marcados con ^{18}O mediante trampa lineal (LTQ).

Con objeto de desarrollar un marco estadístico adecuado para el análisis masivo de cambios de expresión de péptidos marcados con ^{18}O, llevamos a cabo un experimento de hipótesis nula a gran escala. En este experimento se prepararon dos alícuotas idénticas de 1 mg de extracto de proteínas solubles de células HUVEC para ser comparadas una frente a la otra, y las muestras se digirieron, marcaron con ^{18}O y fraccionaron por intercambio iónico, como se detalla en Material y Métodos. Se obtuvieron más de 13,000 PSMs validados al 5% FDR con el método de la razón de probabilidades, que correspondieron a más de 4,800 péptidos y más de 2,400 proteínas diferentes (para mayor detalle consultar la Tabla R. 1). La cuantificación se efectuó sobre los espectros ZoomScan correspondientes a los péptidos identificados mediante un algoritmo previamente desarrollado en el laboratorio (A Ramos-Fernandez et al., 2007) que en este trabajo hemos implementado en el programa QuiXoT (véase capítulo *Desarrollo de un software de cuantificación para marcaje isotópico con ^{18}O*). QuiXoT permite el control de la eficiencia de marcaje de todos y cada uno de los péptidos cuantificados, mediante un ajuste del espectro a una curva teórica definida simultáneamente por las concentraciones de péptido (en unidades de área) de las dos muestras analizadas, la eficiencia de marcaje, y otros factores que definen este tipo de espectros (señal de fondo, ancho de pico y leptocurtosis). Basándose en estos parámetros se realizó un primer filtrado desechando aquellas cuantificaciones que produjeron valores sin sentido físico y que, por tanto, se trataban de malos ajustes (ver Tabla R. 1). También se eliminaron los péptidos no trípticos (pertenecientes al extremo carboxilo de las proteínas), que no pueden ser marcados mediante ^{18}O. Finalmente, también se comprobaron los valores de eficiencia de marcaje de los péptidos, encontrándose que la inmensa mayoría de los péptidos tenían un marcaje superior al 80%, como se aprecia en la Figura R. 9.

	Test	VEGF	
		4 h	8 h
Fracciones de intercambio catiónico	71	64	92
Espectros MS/MS obtenidos	142.585	134.590	149.713
Espectros MS/MS que corresponden a un péptido[a]	13.701	9.171	12.196
Número de péptidos identificados	4.878	4.806	3.876
Número de proteínas identificadas	2.461	1.982	2.085
Espectros restantes tras un filtrado inicial[b]	9.786	7.315	8.145
Espectros utilizados para el análisis[c]	7.640	5.097	4.391
Número de péptidos cuantificados	2.271	2.556	1.218
Número de proteínas cuantificadas	1246	1.278	890
Varianza de espectros (σ^2_S) (95% I.C.[d])	0.018 (0.016-0.020)	0.024 (0.022 - 0.026)	0.031 (0.028 - 0.034)
Varianza de péptidos (σ^2_P) (95% C.I.)	0.021 (0.014-0.028)	0.019 (0.014 - 0.026)	0.014 (0.006 - 0.023)
Varianza de proteínas (σ^2_Q) (95% C.I.)	0.0007 (0-0.009)	0.004 (0 - 0.12)	0.003 (0 - 0.011)
Cambios de expresión estadísticamente significativos[e]	1	26	37

[a] con un FDR igual o menor del 5%, calculado usando el método de la probability ratio.

[b] "Espectros" aquí es referido a espectros tipo ZoomScan. El filtro inicial sigue los siguientes criterios: $\sigma < 0{,}12$, $\beta < 1{,}1$, $f < 1{,}2$ and $v_{qps} > 3{,}1$.

[c] Restantes tras un segundo filtro: eliminación de péptidos C-terminales, péptidos que contuvieran metionina en su secuencia, o sitios de corte perdidos y sus subpéptidos encontrados en la muestra.

[d] I.C., intervalo de confianza.

[e] Desvíos atípicos a nivel de proteína con FDR_q menor del 5% .

Tabla R. 1 Estadísticas de identificación y cuantificación de los tres experimentos a gran escala descritos en el texto

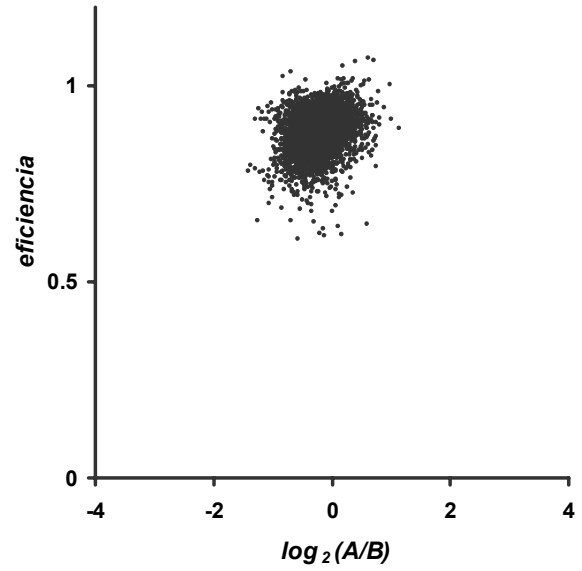


Figura R. 9 Análisis de eficiencia de marcaje. Distribución de las eficiencias de marcaje de los péptidos en función de los ratios en logaritmo en base 2 en el experimento de prueba usado para validar la hipótesis nula.

Análisis mediante métodos estadísticos clásicos.

Para analizar la hipótesis nula, se estudió la distribución de $\log_2$(A/B), donde A y B son las concentraciones de péptido en la muestra no marcada y en la muestra marcada, respectivamente, en unidades de área de los picos en los espectros ZoomScan (Figura R. 10 A). Para evitar posibles artefactos, se excluyeron los péptidos que contuvieran metionina y fueran, por tanto, susceptibles de oxidarse convirtiéndose en una especie diferente, y aquellos péptidos que hubieran sido objeto de una digestión parcial (aquellos péptidos conteniendo sitios trípticos de corte dentro de la secuencia, así como sus respectivos subpéptidos trípticos). El comportamiento de estos dos grupos de péptidos se analiza más adelante.

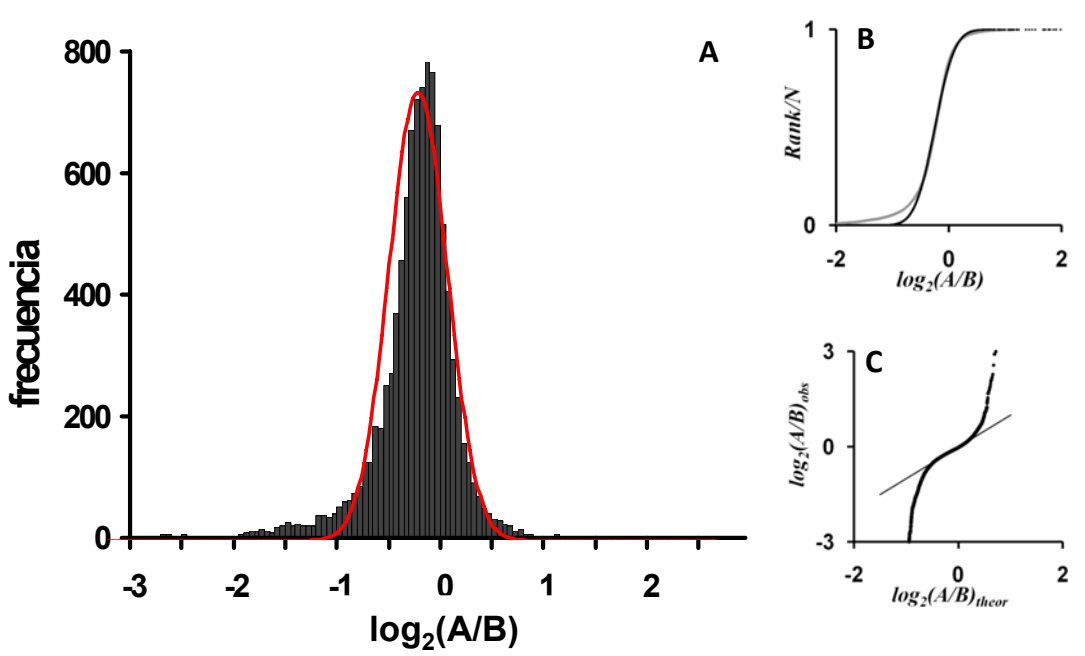


Figura R. 10 Análisis de los resultados del experimento de prueba asumiendo que en la hipótesis nula los datos se distribuyen normalmente. (A) Histograma de frecuencias de los ratios (expresados en logaritmo en base 2), y ajuste del histograma a una distribución gaussiana. (B) distribución acumulativa de frecuencias (en gris), mostrando la curva teórica correspondiente al mejor ajuste de los datos a una distribución normal (en negro). (C) Gráfica de normalidad (la línea delgada representa los resultados esperados para una distribución normal).

Se representó la distribución de $\log_2$(A/B) en un histograma de frecuencias (Figura R. 10 A). La distribución es aparentemente simétrica, siguiendo la forma de una distribución normal. No obstante, al analizar la distribución acumulativa normalizada y su ajuste a una distribución

gaussiana teórica (figura Figura R. 10 B) y la llamada "gráfica de normalidad" (que representa el valor normalizado observado de la variable respecto al valor normalizado teórico y que en casos de distribución normal pura deben producir una recta) (RB D'Agostino et al., 1990) (Figura R. 10 C), se hizo patente una clara desviación del comportamiento normal. De hecho, el comportamiento en los extremos de la función representada en Figura R. 10 C indicaba que la distribución estudiada podría ser fruto de la súper imposición de diversas distribuciones normales con distintas varianzas (RB D'Agostino et al., 1990), cuyo resultado no es una distribución gaussiana.

Con objeto de analizar los resultados con más detalle, se llevó a cabo un test de normalidad (RB D'Agostino, 1971) sobre el conjunto de datos de la muestra, confirmándose que los datos no eran compatibles con una distribución gaussiana. Estos resultados corroboraron que este tipo de datos de cuantificación masiva no pueden describirse usando métodos estadísticos clásicos basados en distribuciones normales. Consistentemente, en este experimento de hipótesis nula se observaron más de 200 péptidos (correspondientes a 72 proteínas) que cambiaron de expresión de forma estadísticamente significativa con un p-value < 0.05, y más de 40 péptidos (20 proteínas) se detectaron como cambios de expresión estadísticamente significativos, si se utiliza como factor discriminante una FDR de cambio de expresión < 5%. Estos resultados son inaceptables desde el punto de vista estadístico, ya que en estas condiciones una gran proporción de los cambios de expresión que se detectarían en experimentos reales usando estos criterios estadísticos serían falsos.

Buscando la normalización: definición de peso estadístico

El indicio de que la distribución global podía ser la suma de varias distribuciones normales con diferente varianza nos hizo reflexionar sobre las posibles causas de esta diversidad de varianzas. Después de una inspección cuidadosa observamos que "la calidad" de cada uno de los espectros puede ser un factor determinante sobre la precisión de la medida. Como se observa en la Figura R. 11, dos espectros del mismo experimento que cuantifican el mismo péptido pueden dar lugar a valores muy diferentes, debido a que uno de ellos es evidentemente de baja calidad y por tanto es menos fiable. En otras palabras, el error cometido en las cuantificaciones depende de la calidad de los espectros y no es, por tanto, constante.

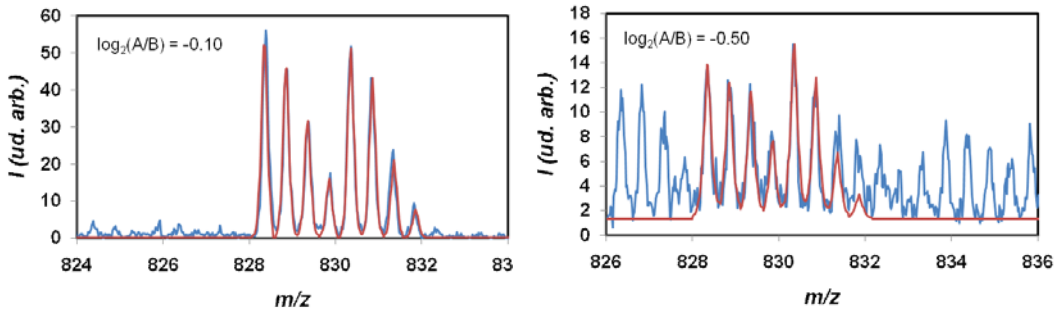


Figura R. 11 Calidad de los espectros. En el panel (A) se observa un espectro cuya una relación señal/ruido es alta, en el que la cuantificación es muy precisa. En (B) se observa una cuantificación del mismo péptido en un espectro con una relación señal/ruido muy inferior, en el que la cuantificación es mucho menos precisa y por tanto se lleva a cabo con mayor margen de error . Nótese que los valores de cuantificación difieren notablemente.

Nos planteamos resolver este problema clasificando los espectros en función de cierto parámetro de calidad, de modo que a cada una de las categorías se pudiera asignar un error determinado – o varianza. Si la clasificación es suficientemente buena, los análisis de normalidad efectuados sobre cada una de estas categorías deberían dar un resultado positivo. En una primera aproximación se pensó en clasificar los espectros por su intensidad, ya que intuitivamente cabe esperar que los espectros más intensos den lugar a resultados más fiables. Como se observa en la Figura R. 12, este criterio de intensidad parece generar un orden adecuado – a mayor intensidad la dispersión parece menor –. Sin embargo, también se observó que aún se encontraban muchos espectros que se desviaban atípicamente del valor esperado de acuerdo a su intensidad, y los test de normalidad aplicados localmente sobre categorías de espectros teniendo una intensidad semejante no dieron resultado positivo. Al estudiar más detenidamente los espectros que se desviaban, se observó que muchos de ellos eran espectros cuya curva teórica no estaba bien ajustada, bien por un error en la asignación del péptido, que hace que la masa del ión precursor no se aproxime a la masa calculada según la secuencia del péptido, o bien por la presencia de otra especie co-eluída con una relación carga-masa similar, que deformaba la envoltura isotópico del péptido identificado alterando su cuantificación.

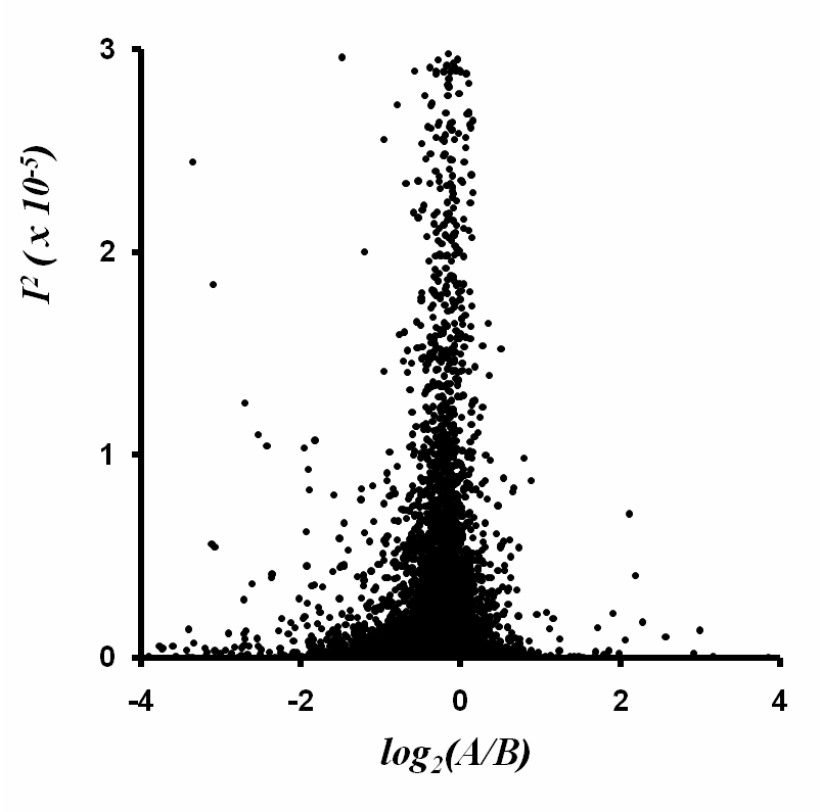


Figura R. 12 Peso estadístico basado en la intensidad. Distribución del cuadrado de las intensidades en función del $\log_2$(ratio).

Se decidió entonces probar un factor de ordenación basado exclusivamente en el error cometido al realizar el ajuste de la curva teórica – que puede determinarse a partir de los errores de ajuste de cada uno de los parámetros – utilizando la teoría de propagación de errores. Si se conoce el error cometido en la concentración de péptido en las dos muestras, A y B, en unidades de área, el error cometido al calcular el logaritmo de la razón respectiva es:

$$si \quad f = \log_2(A/B),$$
$$\sigma_f^2 = \left(\frac{\partial f}{\partial A}\right)^2 \sigma_A^2 + \left(\frac{\partial f}{\partial B}\right)^2 \sigma_B^2 = \log_2(e)\left(\frac{\sigma_A^2}{A^2} + \frac{\sigma_B^2}{B^2}\right)$$

Ecuación R. 10

Y como el peso estadístico en la media ponderada de varias magnitudes con diferente error es la inversa de su varianza, podemos definir un peso $v_{A,B}$:

$$v_{A,B} = \frac{1}{\log_2(e)\left(\frac{\sigma_A^2}{A^2} + \frac{\sigma_B^2}{B^2}\right)}$$

Ecuación R. 11

Nótese que este nuevo peso, deducido teóricamente, es adimensional, aumenta con el cuadrado de la intensidad de la medida (reflejando así el efecto intensidad) y disminuye con su desviación cuadrática (reflejando así el efecto de la bondad de ajuste a la curva teórica).

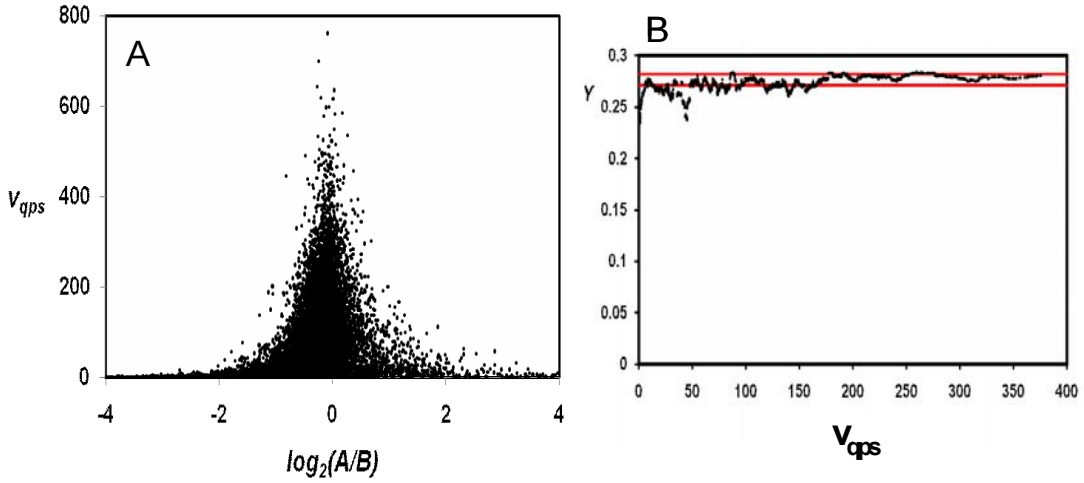


Figura R. 13 Peso estadístico basado en la teoría de propagación de errores. (A) Distribución del peso estadístico, definido por la ecuación R.11, en función del log_2(ratio). (B) Valores del estadístico Y del test de D'Agostino calculado de forma local en función del peso estadístico; este test se calcula en poblaciones deslizantes que contienen las 200 medidas que tienen un peso estadístico más próximo. En rojo se marcan los límites de confianza del estadístico Y en distribuciones conteniendo 200 valores.

Al representar los datos clasificados por este peso estadístico en la Figura R. 13 A, observamos que la distribución sigue ahora un orden muy coherente, generando una nube de puntos cuya desviación va disminuyendo a medida que aumenta el peso, y además no se observan desviaciones atípicas para prácticamente ningún valor del peso estadístico. Además, los test de normalidad aplicados a poblaciones generadas usando una ventana deslizante de doscientas cuantificaciones en torno al peso más próximo, dieron un resultado positivo en todos los casos (Figura R. 13 B), excepto para pesos estadísticos menores a un valor umbral de peso de 3.1. Este umbral límite, que separa los valores de tan baja calidad que por debajo de

este peso resulta imposible modelar la dispersión utilizando una distribución normal, se empleó para eliminar de forma coherente las cuantificaciones de calidad insuficiente.

Sin embargo, al utilizar el nuevo peso en análisis posteriores observamos que al analizar experimentos de proteómica cuantitativa real – y no de hipótesis nula – los cambios de expresión se atenúan, como se puede apreciar en el ejemplo de un proteoma analizado en la Figura R. 14 A y B. Esta atenuación se debe a que cuando tiene lugar un cambio de expresión, una de las cantidades, A ó B, tiene un valor pequeño y, por tanto, su error relativo aumenta, de manera que hay una correlación negativa entre el peso y el cambio de expresión. Por otro lado observamos que muchos espectros claramente influenciados en su cuantificación por la presencia de un co-eluido presentaban un peso estadístico anormalmente alto.

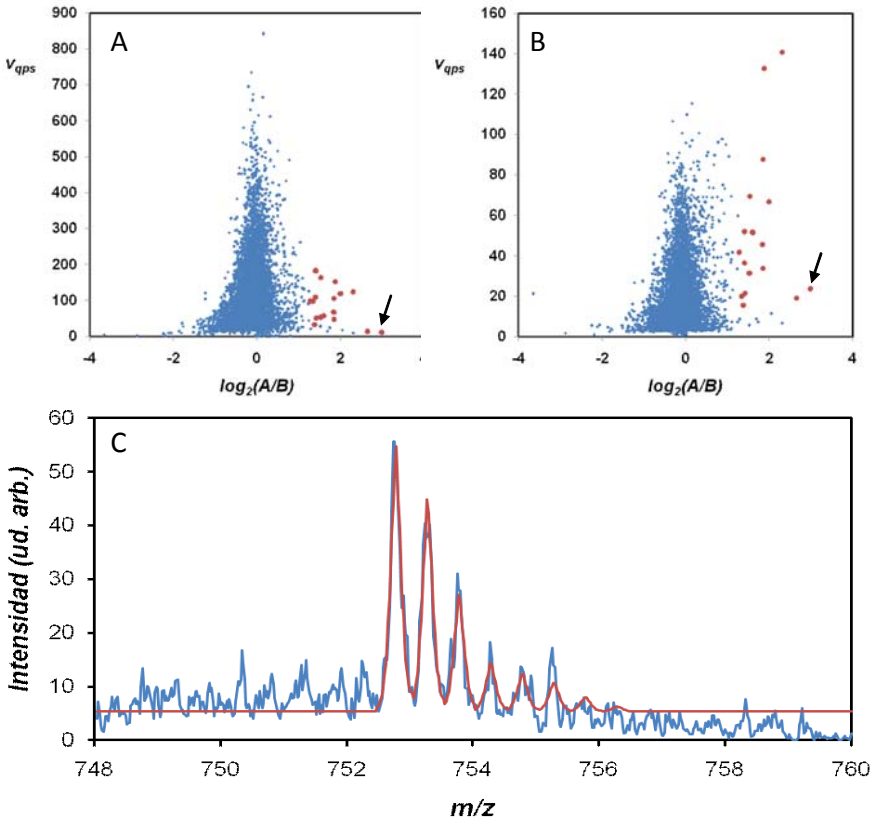


Figura R. 14 Comparación de pesos estadísticos. (A) Distribución de pesos estadísticos definidos por la ecuación R.11 en un proteoma con cambios de expresión (marcados en rojo). (B) Distribución de pesos estadísticos basados en la ecuación R.12 del mismo experimento, donde se destacan los mismos cambios. (C) Espectro de cuantificación correspondiente a uno de los cambios de expresión detectados en este experimento (indicado por una flecha en los paneles A y B).

Para corregir estos efectos se ideó un nuevo peso estadístico, basado en la desviación cuadrática media entre el espectro experimental y su curva teórica, un valor que, al representar la varianza global de la función ajustada, es más robusto que la desviación cuadrática de A y de B por separado (nótese que estos dos valores se calculan a partir de esa misma magnitud). La desviación cuadrática media de la función se calculó en la región del espectro que contiene la envoltura isotópica de la especie no marcada y de la marcada, en un rango de 8 Th, así como en dos regiones laterales de 2 Th que evalúan desviaciones debidas a la presencia de especies co-eluyentes próximas (véase Figura R. 15). El peso estadístico de una medida (espectro) *s*, proveniente de la cuantificación del péptido *p* perteneciente a la proteína *q*, se determinó, finalmente, como

$$v_{qps} = \frac{T^2}{MSD_C + MSD_L}$$

Ecuación R. 12

donde *T* es una medida de la concentración de péptido de la especie más intensa en unidades de área, MSD_C es la desviación cuadrática media en la región conteniendo las dos envolturas isotópicas del péptido y MSD_L es la desviación cuadrática media en uno de los laterales, escogido en función de la especie del péptido más intensa. Si $A+B_0 > B_1 + B_2$, (siendo A, B_0, B_1, y B_2 las cantidades de péptido estimadas para la especie no marcada, la especie marcada en la que no se incorporó ningún ^{18}O, la especie marcada en la que se incorporó solamente un ^{18}O y la especie marcada en la que se incorporaron dos ^{18}O, tal y como se explica en trabajos previos(A Ramos-Fernandez et al., 2007)), MSD_L se calcula en la región izquierda del péptido y $T = A$. Si $A+B_0 < B_1+B_2$ entonces MSD_L se calcula en la región derecha, y $T = B$.

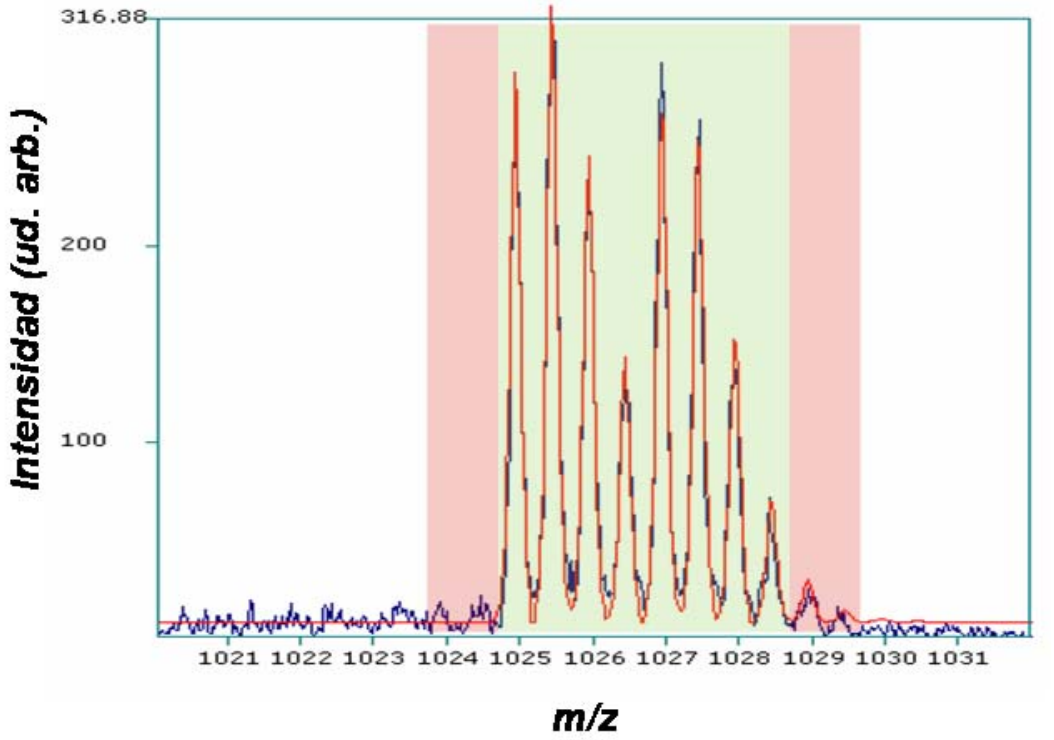


Figura R. 15 Esquema mostrando las regiones del espectro en las que se evalúa la desviación cuadrática media. En verde se marca la región considerada como la envoltura isotópica del péptido (MSD_C), y en rojo las regiones laterales (MSD_L) que se toman en cuenta en el peso estadístico definido en la Ecuación R. 12.

Este nuevo peso estadístico permite una clasificación eficaz de las cuantificaciones (Figura R. 16), que verifica el test de normalidad local (exceptuando los pesos por debajo del umbral mínimo), como se observa en Figura R. 16 A-interior. A modo de ejemplo, si se escoge cualquier rango de pesos, como el marcado en Figura R. 16 A, se comprueba que las cuantificaciones en dicha región se comportan de acuerdo a una distribución normal, como indican las Figura R. 16 B y C.

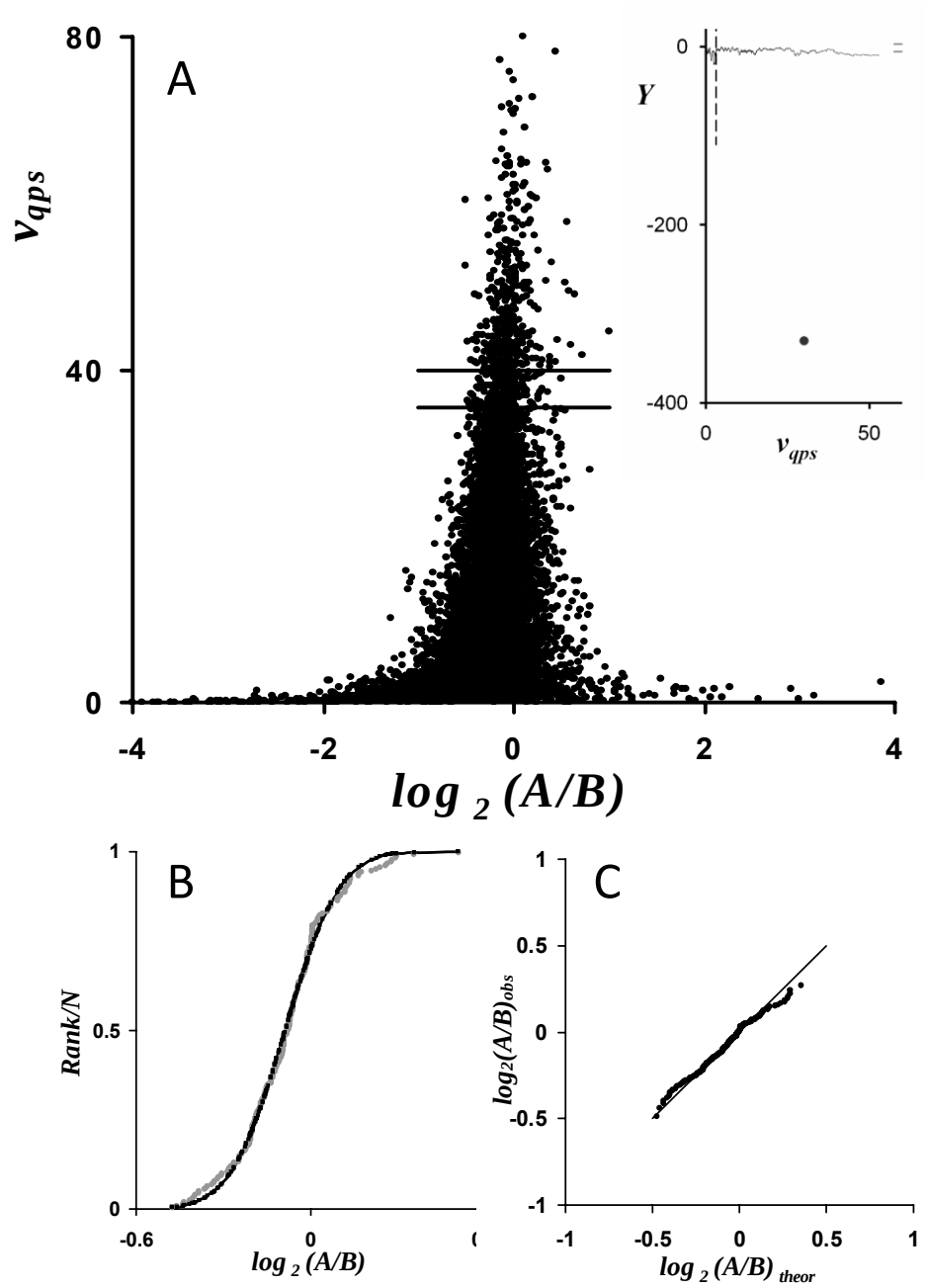


Figura R. 16 Estratificación de las medidas de cuantificación diferencial en poblaciones localmente normales de acuerdo a su peso estadístico. (A) Distribución de los pesos estadísticos calculados según la ecuación R.12 en función de los log2(ratio). (A,interior) Resultados del test de normalidad de D'Agostino aplicado en ventanas deslizantes de 200 puntos ordenados por el peso estadístico; se muestran los valores del estadístico Y del test de D'Agostino, indicando los valores del 99% de confianza para una distribución normal de 200 datos (2.9 a -5.6); compárense estos valores con el valor de Y obtenido para el conjunto total de datos (indicado por un punto negro). (B) y (C) Gráficas de normalidad de una de las regiones de (A) tomada a modo de ejemplo (delimitada por barras horizontales).

Un modelo con tres niveles de varianza

Teoría general

Además del error cometido en la medida, y que se modeló satisfactoriamente según se describe en el apartado anterior, existen también otras dos fuentes de error debidas, por un lado, a la cuantificación de las proteínas a partir de sus péptidos respectivos y, por otro, a la variabilidad en la concentración de proteína introducida durante la manipulación de la muestra. Las tres fuentes de error se integraron utilizando un modelo de efectos aleatorios (*random effects model*) , que se define como se describe a continuación. Sea x_{qps} el logaritmo en base dos de la relación entre las concentraciones *A* y *B* del péptido marcado y no marcado determinada por un espectro *s* correspondiente al péptido *p* que proviene de la proteína *q*, es decir:

$$x_{qps} = \log_2(A/B)$$

Ecuación R. 13

Se asume que los errores experimentales de esta medida provienen de un error sistemático, μ, en la razón en la que están mezcladas las dos muestras; de las desviaciones que se producen en la concentración de proteína debidas a la variabilidad biológica y durante el proceso de preparación de los extractos de proteína ρ_q; de desviaciones en la concentración de péptido generado por digestión tríptica de su proteína correspondiente, β_{qp}; y del error de cuantificación de la pareja peptídica, ξ_{qps} , que puede ser debido por una parte al error de medida inherente al espectrómetro de masas utilizado y por otra al ajuste del espectro experimental a una curva teórica (evaluado por el peso estadístico v_{qps} según se ha descrito en el apartado anterior), es decir:

$$x_{qps} = \mu + \rho_q + \beta_{qp} + \xi_{qps}$$

Ecuación R. 14

Se asume que β_{qp} y ρ_q siguen distribuciones normales, es decir, $\beta_{qp} \sim N(0,\sigma^2_P)$ y $\rho_q \sim N(0,\sigma^2_Q)$, y que las varianzas de péptido y proteína, σ^2_P y σ^2_Q respectivamente, son constantes. También se asume que ξ_{qps} se distribuye normalmente como $\xi_{qps} \sim N(0,\sigma^2_S + k/v_{qps})$, donde σ^2_S

es la varianza de la cuantificación de los espectros debida al error de medida del espectrómetro de masas, v_{qps} es el peso estadístico asociado al ajuste a una curva teórica y *k* es una constante de proporcionalidad para el peso estadístico que depende de la definición que se ha escogido para el peso estadístico y del espectrómetro de masas y de las condiciones en las que se ha programado el aparato para adquirir los espectros.

La constante *k* se calcula ordenando las cuantificaciones por su peso estadístico y representando la desviación cuadrática media (*MSD*) de la medida respecto al péptido correspondiente en una ventana deslizante conteniendo un número suficientemente alto de medidas (normalmente 200), en función del peso estadístico v_{qps}, y ajustando la curva obtenida a la función:

$$MSD = \sigma_S^2 + k/v_{qps}$$

Ecuación R. 15

Como la cuantificación de cada proteína se lleva a cabo por varios péptidos, y cada péptido a su vez se cuantifica por varios espectros, se escoge una media ponderada para cada uno de estos valores, donde el peso estadístico *w* a nivel de espectro *s*, péptido *p* y proteína *q* es la inversa de sus varianza local, y la varianza local a cada nivel se calcula, de acuerdo a la teoría estadística, como la inversa de la suma de las inversas de las varianzas de las medidas en el nivel anterior, es decir

$$w_{qps} = \frac{1}{\frac{k}{v_{qps}} + \sigma_S^2} \quad ; \quad w_{qp} = \frac{1}{\frac{1}{\sum_s w_{qps}} + \sigma_P^2} \quad ; \quad w_q = \frac{1}{\frac{1}{\sum_q w_{qp}} + \sigma_Q^2}$$

Ecuación R. 16

Nótese que ésta es la forma general de integrar medidas con diferente varianza, y que cuando hay *n* medidas con el mismo error la varianza final se reduce a la conocida fórmula de dividir entre *n* la varianza del nivel anterior.

De manera que las medias ponderadas para péptido, proteína, y media global – que es una estimación del valor μ – se calculan, respectivamente, de la siguiente manera

$$x_{qp} = \frac{\sum_s w_{qps} \cdot x_{qps}}{\sum_s w_{qps}} \quad ; \quad x_q = \frac{\sum_s w_{qps} \cdot x_{qps}}{\sum_s w_{qps}} \quad ; \quad x = \frac{\sum_q w_q \cdot x_q}{\sum_q w_q}$$

Ecuación R. 17

El cálculo de las varianzas es complejo y se lleva a cabo mediante un algoritmo que se describe en el apéndice A. Es importante indicar aquí que no hemos encontrado en la bibliografía ningún caso donde se haya aplicado previamente un modelo estadístico con esta estructura y que el modelo no tiene solución analítica exacta; por eso el algoritmo se basa en un método que se aplica de forma iterativa hasta que se obtiene una convergencia satisfactoria de valores. También es importante comentar que el modelo estadístico dado por las fórmulas anteriores ha sido extensivamente validado mediante simulación por el método de Monte Carlo, utilizando varianzas predefinidas de prueba. Finalmente, el método de Monte Carlo también ha sido utilizado para comprobar que la estima de varianzas realizadas por el método iterativo da resultados no sesgados, lo que resulta particularmente crítico en este tipo de experimentos, donde el número medio de medidas por péptido y el número de péptidos por proteína es un valor próximo a la unidad y por tanto la corrección por grados de libertad, si no se aplica adecuadamente, produce notables errores de sesgo.

Los intervalos de confianza de las varianzas estimadas también se determinan mediante simulación por Monte Carlo. Esta simulación se realiza generando conjuntos de datos aleatorios que contienen árboles completos de espectros/péptidos/proteínas con el mismo tamaño y distribución de número de espectros por péptido y número de péptidos por proteína (**Error! Reference source not found.**) que el conjunto de datos original. Los valores se generan de forma aleatoria con un error que sigue una distribución normal en concordancia con las varianzas calculadas en el experimento real. Generalmente se calculan 100 “experimentos” completos aleatorios para determinar los intervalos de confianza. Se muestra un ejemplo en las Figura R. 18 y Figura R. 17.

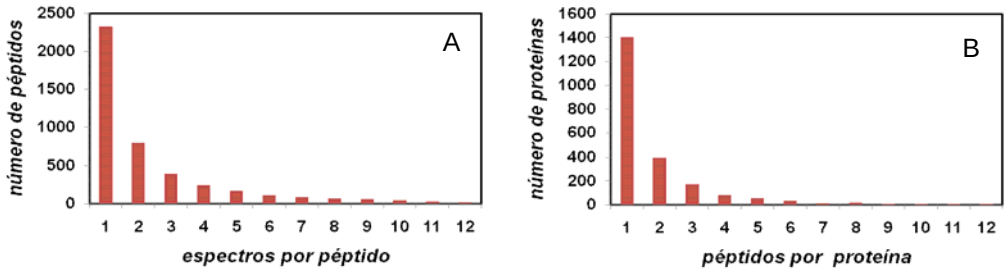


Figura R. 18 Distribuciónes del número de medidas por péptido y del número de péptidos por proteína típicos en un experimento de proteómica cuantitativa. Estas distribuciones corresponden a los resultados del experimento de prueba de la hipótesis nula.

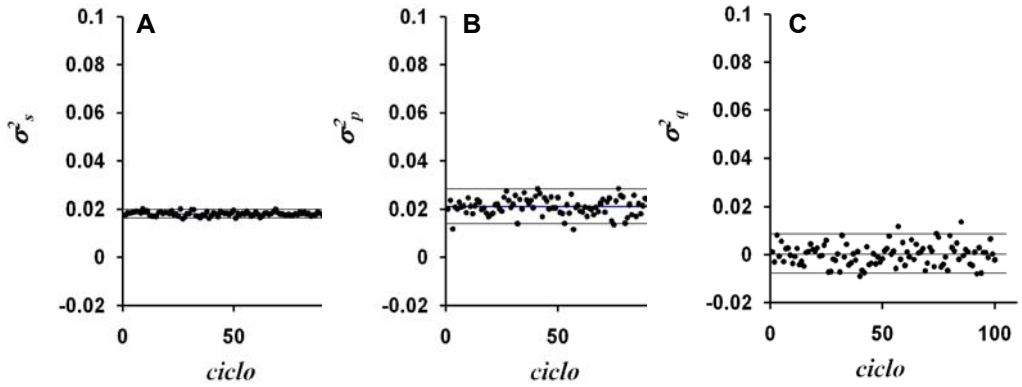


Figura R. 17 Cálculo de los intervalos de confianza de las varianzas estimadas mediante simulación de Monte Carlo. Se simulan 100 experimentos con el mismo número de proteínas que el experimento analizado, siguiendo la distribución de péptidos por proteína y medidas por péptido de la figura R.16 y generando errores aleatorios a nivel de medida, de péptido y de proteína, distribuidos normalmente de acuerdo a las varianzas estimadas a partir de los datos experimentales; las varianzas se vuelven entonces a estimar a partir de los datos simulados y el conjunto de varianzas obtenidas en las 100 simulaciones se utiliza para determinar los intervalos de confianza del 95%. En la gráfica se representan las varianzas σ^2_S (A), σ^2_P (B) y σ^2_Q (C). Las barras horizontales muestran el intervalo de confianza del 95%, y el valor estimado en el experimento real.

Detección semiautomática de valores atípicos (outliers) utilizando el modelo estadístico

El modelo estadístico permite determinar cuál es la varianza local de cada una de las medidas o de cada uno de los péptidos o proteínas, que viene dada por la inversa de los respectivos pesos, calculados de acuerdo a las Ecuación R. 16. Estas varianzas locales pueden utilizarse para estimar si un valor determinado – un espectro, un péptido o una proteína – se desvía significativamente de la media correspondiente – el péptido, la proteína o la media global, respectivamente – ya que de acuerdo al modelo, a todos los niveles las medidas se comportan como distribuciones normales locales.

Si $P(\mu,\sigma^2_{xt})$ es la probabilidad de que el valor x_t correspondiente a la medida *t* se desvíe de la distribución normal con media μ y varianza σ^2, la probabilidad de que una medida a nivel de espectro, péptido o proteína se desvíe de la media correspondiente viene dada por

$$p_{qps} = P\left(x_{qp}, w_{qps}^{-1}, x_{qps}\right) \quad ; \quad p_{qp} = P\left(x_q, w_{qp}^{-1}, x_{qp}\right) \quad ; \quad p_q = P\left(x, w_q^{-1}, x_q\right)$$

Ecuación R. 18

La presencia de valores atípicos se detecta utilizando una prueba de hipótesis múltiple y controlando la tasa de error FDR, definida como la proporción de valores que se espera que se desvíen de la distribución normal al azar dentro de la población de valores atípicos:

$$FDR_{qps} = \frac{p_{qps} \cdot NS}{O\left(p_{qps}\right)} \quad ; \quad FDR_{qp} = \frac{p_{qp} \cdot NP}{O\left(p_{qp}\right)} \quad ; \quad FDR_q = \frac{p_q \cdot NQ}{O\left(p_q\right)}$$

Ecuación R. 19

donde los numeradores evalúan el número esperado de eventos (la probabilidad calculada anteriormente multiplicada por el número de casos evaluado), y el denominador es el número de eventos observados con una probabilidad igual o menor. De esta forma, si no existen valores atípicos, y la distribución estudiada sigue correctamente las propiedades de una distribución normal, todos los valores de FDR serán próximos a uno, y un valor atípico será detectado por tener un valor de la FDR correspondiente próximo a cero.

Análisis de valores atípicos a nivel de espectro y de péptido

En la Figura R. 19 se observan los valores atípicos encontrados para tres experimentos, de los cuales uno de ellos (Figura R. 19 A) es el experimento de hipótesis nula. Tras una eliminación previa de aquellos espectros que debido a un mal ajuste -- un co-eluído afectando a la cuantificación o un bajo nivel señal-ruido – el valor de la cuantificación difiere claramente de la media de su péptido correspondiente, puede verse que el número de valores atípicos encontrados a nivel de la medida del espectro es muy bajo, lo que demuestra la robustez del método utilizado.

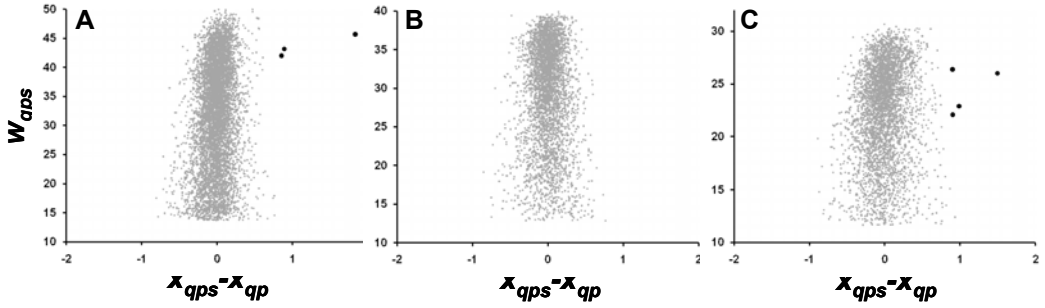


Figura R. 19 Análisis de cuantificaciones atípicas a nivel de medida. Se representa el peso estadístico en función de la diferencia entre el valor de la medida y el promedio a nivel de péptido. Los resultados corresponden al experimentos de prueba (A) y a los experimentos de estimulación de células HUVEC por VEGF durante 4h (B) y 8h (C). Los puntos negros indican cuantificaciones atípicas a nivel de medida (FDR_{qps} < 5%).

En el caso de los péptidos, se detectaron numerosísimos casos con valores atípicos que indujeron a pensar en un problema añadido. En varios de los experimentos realizados, como se ve en la **Figura R. 20** hay un claro artefacto debido a la oxidación parcial de péptidos con metionina en su secuencia, que al no tener lugar en la misma extensión en las dos muestras introduce variabilidad en la cuantificación. De la misma manera afecta la digestión parcial de ciertos péptidos, que puede no tener lugar de la misma manera en las dos muestras. Este tipo de artefactos aumentan la varianza a nivel de péptido $-\sigma^2_P$ – y el número de valores atípicos que se detectan por debajo de cierto umbral de FDR_{qp}. Estos casos son un buen ejemplo de la utilidad del análisis de las FDR y de las varianzas a los tres niveles, que permiten detectar errores causados por un defecto en el protocolo (ya sea en la extracción, en la digestión o en el

marcaje), o algún error de ejecución en el mismo(E Bonzon-Kulichenko et al., en revisión). Otro ejemplo de problemas detectados en un protocolo de marcaje a nivel de péptido es el mostrado en la Figura R. 21, donde se observan las consecuencias de un problema de sobrealquilación en una muestra marcada.

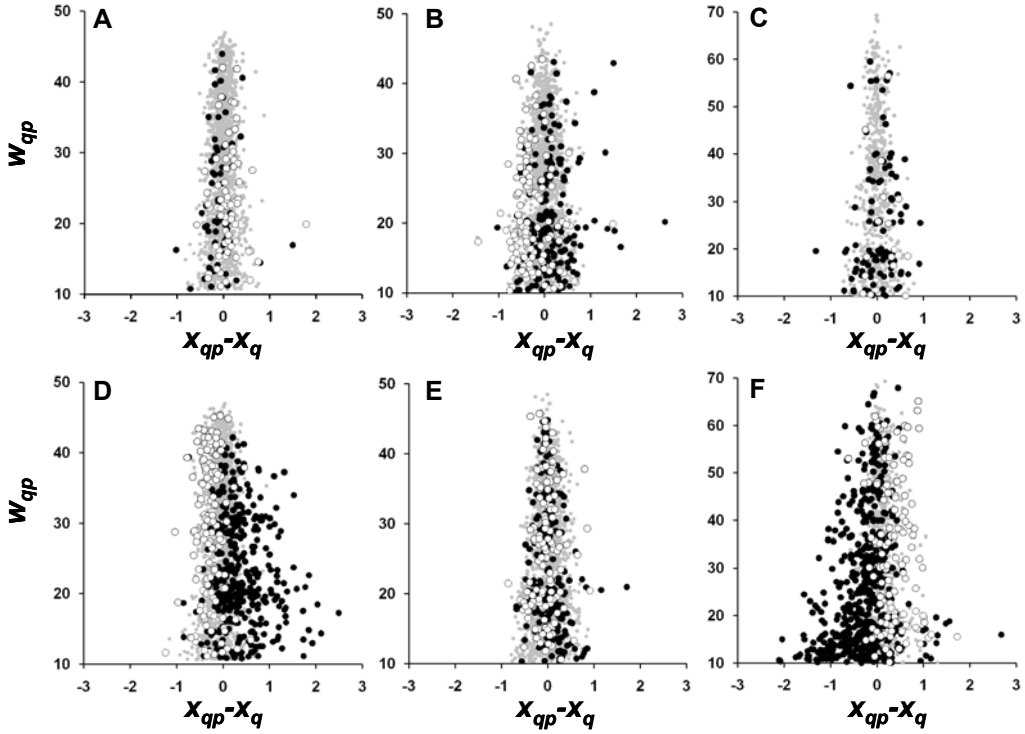


Figura R. 20 Análisis de cuantificaciones atípicas a nivel de péptido (I). En estas gráficas se analiza el efecto de la oxidación de metioninas (A,B y C), y de las digestiones parciales (D,E y F). Se representa el peso a nivel de péptido en función de la diferencia entre la medida a nivel de péptido y el promedio a nivel de proteína. Se muestran los resultados correspondientes al experimento de prueba (A,D), y a la estimulación de HUVEC con VEGF durante 4h (B,E) y 8h (C,F). En los paneles A, B y C se destacan los péptidos con metionina no oxidada (negro) y aquellos con alguna metionina oxidada (blanco). En los paneles D, E y F se destacan los péptidos de digestión parcial (negro) y sus correspondientes subpéptidos (blanco).

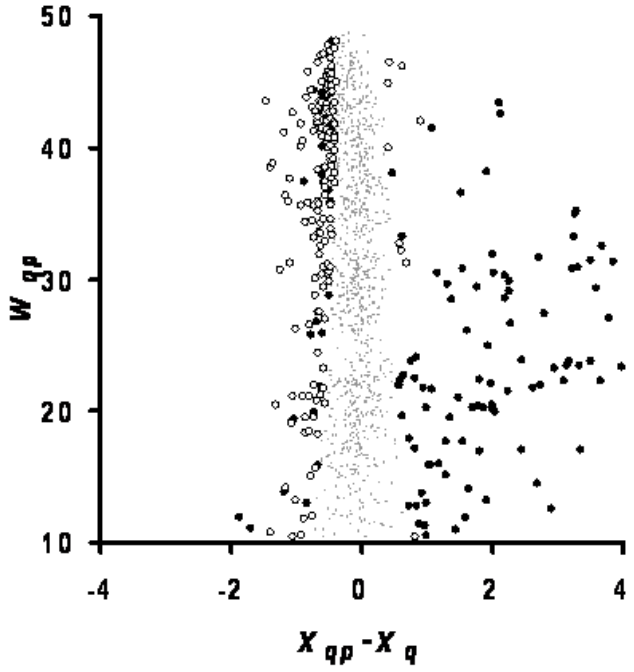


Figura R. 21 Efecto de la sobrealquilación debida al doble tratamiento con iodoacetamida durante el proceso de marcaje con O^{18} de un proteoma. En el análisis de esta muestra se detectó un fenómeno no deseado de alquilación de residuos Lys o grupos amino N-terminales que da lugar a artefactos a la hora de cuantificar los péptidos. En negro se destacan los péptidos alquilados en N, y en blanco los correspondientes péptidos no alquilados.

Una vez se eliminaron del estudio las poblaciones de péptidos conteniendo residuos de metionina o mostrando digestión parcial, puede verse (Figura R. 22) que el número de valores atípicos detectados a nivel de péptido es muy reducido (alrededor de cuatro casos por experimento, en experimentos con más de 4,000 péptidos).

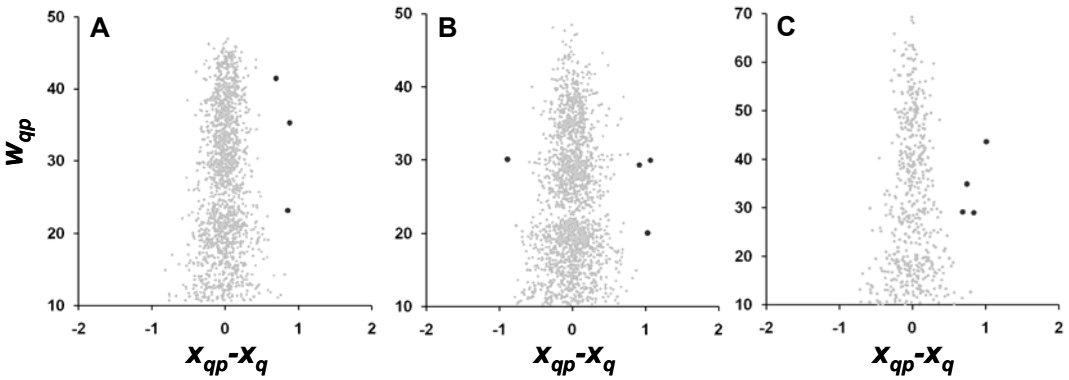


Figura R. 22 Análisis de cuantificaciones atípicas a nivel de péptido (II). Se representa el peso a nivel de péptido en función de la diferencia entre la medida a nivel de péptido y el promedio a nivel de proteína. Se muestran los resultados correspondientes al experimento de prueba (A,D), y a la estimulación de HUVEC con VEGF durante 4h (B,E) y 8h (C,F). Los puntos negros indican las cuantificaciones atípicas a nivel de péptido. (FDRqp < 5%).

Comprobación de la hipótesis nula a nivel de proteína.

Desde un punto de vista meramente estadístico, los valores atípicos que se observan en una distribución de cuantificaciones de proteínas corresponden a los cambios de expresión estadísticamente significativos, y se detectan utilizando la misma estrategia de detección de valores atípicos que se ha descrito anteriormente (Ecuación R. 19). En el experimento de hipótesis nula propuesto, sin embargo, los valores atípicos detectados no pueden ser cambios de expresión, sino que tienen que ser necesariamente artefactos del método. Se inspeccionó en este experimento de hipótesis nula (representado en la Figura R. 23) las cuantificaciones de proteína con $FDR_q < 0.01$, y solamente se encontró un valor atípico de entre las más de 1,200 proteínas cuantificadas en el experimento (ver Tabla R. 1), indicando que la descripción de la hipótesis nula dada por nuestro modelo es correcta.

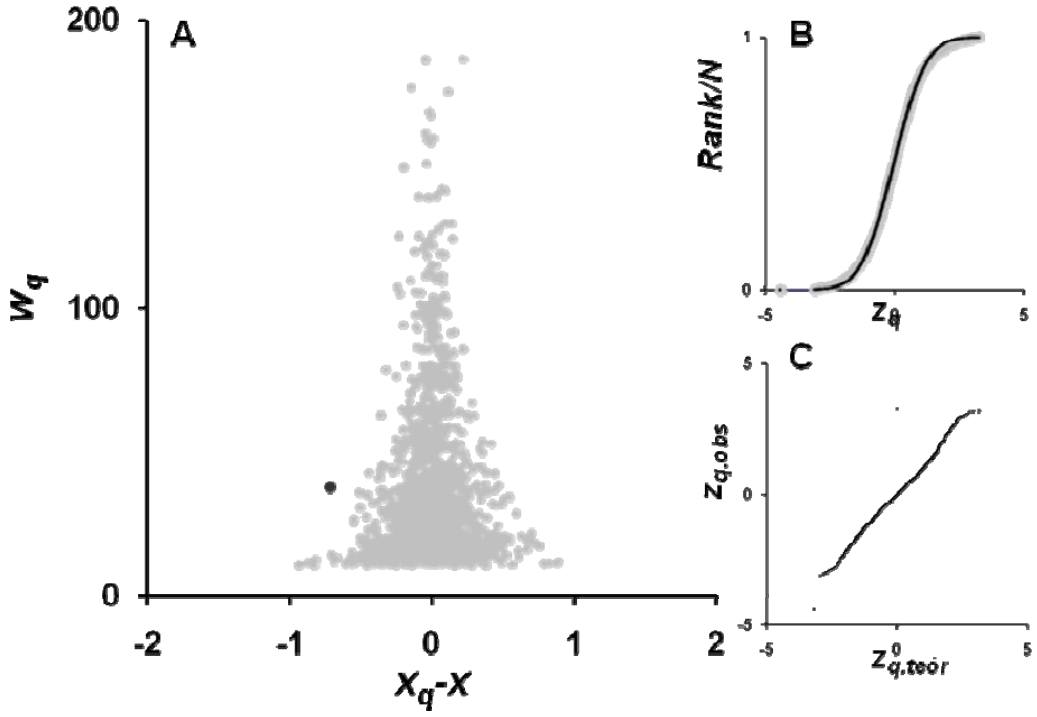


Figura R. 23 Comprobación de la hipótesis nula a nivel de proteína. (A) Representación del peso estadístico en función del log_2(ratio) a nivel de proteína obtenidos en el experimento de prueba. En negro se destaca el único cambio de expresión falso detectado (FDR_q<5%). (B) Distribución acumulativa normalizada de frecuencias de x_q frente a la variable estandarizada z_q; la curva negra representa la distribución teórica

La validación de la hipótesis nula puede realizarse también estudiando la distribución de las cuantificaciones en función de la varianza estimada para cada una de las proteínas. Para analizar conjuntamente estas cuantificaciones, y puesto que cada una de ellas tiene una varianza diferente, las cuantificaciones de proteína x_q se pueden normalizar restando la media global y dividiendo entre su desviación estándar (la raíz cuadrada de la varianza, que a su vez es la inversa del peso estadístico a nivel de proteína), es decir:

$$z_q = \left(x_q - x\right) \cdot \sqrt{w_q}$$

Ecuación R. 20

donde z_q es la medida normalizada a nivel de proteína. Si el modelo estadístico es correcto y las varianzas están bien calculadas, la distribución de z_q debe ser normal con media cero y varianza uno.

La Figura R. 23 B muestra que la distribución acumulativa de frecuencias de la variable normalizada z_q se puede explicar satisfactoriamente como una distribución normal $z_q \sim N(0,1)$, y su correspondiente gráfica de normalidad (RB D'Agostino et al., 1990) (Figura R. 23 C) no muestra aparentemente ningún desvío de la distribución normal esperada. Estos resultados confirman de manera inequívoca que el modelo estadístico constituye una excelente descripción de la hipótesis nula en este tipo de experimentos.

Detección de cambios de expresión.

El modelo estadístico se probó finalmente en la práctica aplicándolo a la detección de cambios de expresión estadísticamente significativos en experimentos reales. Para ello se utilizó el mismo modelo celular utilizado para el análisis de la hipótesis nula. Las células HUVEC se incubaron en ausencia o presencia del factor de crecimiento VEGF durante cuatro y ocho horas, y se compararon los dos proteomas con la condición basal (sin factor de crecimiento VEGF). Se prepararon alícuotas de 1 mg de proteína por cada una de las condiciones comparadas en estos dos experimentos, en la misma forma que se realizó en el experimento de hipótesis nula. En ambos proteomas se detectaron cambios de expresión, tanto de sobre-expresión de proteínas, como de disminución de éstas (Figura R. 25 A y B). Para probar la coherencia de los cambios de expresión detectados, se realizó un experimento réplica del experimento de incubación de cuatro horas a pequeña escala (se utilizaron 100 µg de proteína por cada uno de los estados a comparar). Este experimento mostró que la tendencia de los cambios era la misma que en el experimento a mayor escala (Figura R. 24). Además, algunos de los cambios de expresión detectados en los dos experimentos de gran escala (cuatro y ocho horas) fueron confirmados por Western-Blot (Pablo Martínez-Acedo, tesis doctoral, no se muestra), validando la correcta identificación de los cambios detectados por el modelo estadístico. El modelo ha sido probado, además, en más de una veintena de proteomas diferentes (algunos de ellos pendientes de publicación (E Bonzon-Kulichenko et al., en revisión)), y en todos los casos se ha comprobado que la distribución de cuantificaciones normalizadas a nivel de proteína es gaussiana.

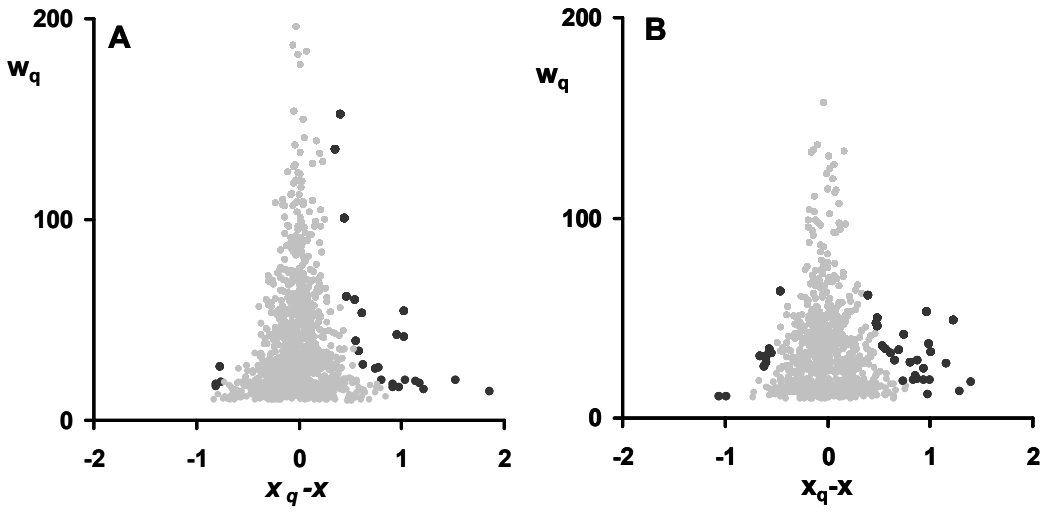


Figura R. 25 Detección de cambios de expresión utilizando el modelo estadístico. Representación del peso estadístico en función del log2(ratio) a nivel de proteína correspondientes a los experimentos de tratamiento de HUVEC con VEGF durante 4h (A) y 8h (B). Los cambios de expresión se detectan utilizando como criterio aquellas desviaciones atípicas a nivel de proteína con FDR_q<5%.

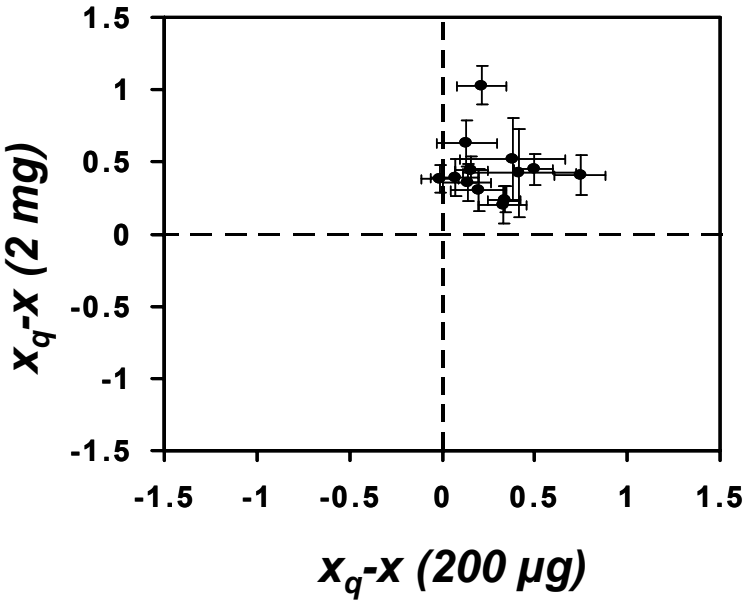


Figura R. 24 Consistencia de los cambios de expresión en los dos experimentos de VEGF 4h. La lista de proteínas que muestra un cambio de expresión significativo a un nivel de FDR_q < 35% (p_q < 0.02) en el experimento VEGF 4h a gran escala es comparada con las mismas cuantificadas en el experimento de réplica a pequeña escala, y los valores de xq de un experimento son representados frente a los observados en el otro. Las barras de error representan las desviaciones estándar de las medias de las cuantificaciones de proteína ($\sqrt{w_q^{-1}}$).

2.2 Generalización del modelo a otros métodos de marcaje mediante isótopos estables (SILAC y iTRAQ) y otros espectrómetros de masas.

Estrategia experimental

Para demostrar que el modelo estadístico puede ser aplicado a otros métodos de marcaje (SILAC y iTRAQ) y otros espectrómetros de masas, se eligió el proteoma de un cultivo de levadura como modelo, según se detalla en Material y Métodos. La levadura tiene la ventaja de poderse cultivar a una escala suficientemente grande para obtener la cantidad de proteína necesaria; por otra parte, se trata de un modelo celular que resulta relativamente fácil de marcar mediante la técnica SILAC. Como estímulo se eligió un tratamiento con H_2O_2 a una concentración (5 mM) que, de acuerdo a un estudio tentativo previo (no se muestra), produce muy pocos cambios de expresión y, por tanto, la inmensa mayoría de las proteínas no experimenta variaciones en su concentración.

Se prepararon dos únicos cultivos celulares de levadura idénticos, uno de ellos en un medio de cultivo convencional (cultivo A) y el otro en un medio con lisina y arginina pesadas (cultivo A*). En un momento dado del proceso, se separó el cultivo A en dos alícuotas idénticas, y a una de ellas se le adicionó H_2O_2 (cultivo B). Las muestras A, A* y B fueron lisadas, y los extractos de proteínas obtenidos se alicuotearon con objeto de poder realizar una batería de experimentos comparativos idénticos en paralelo, utilizando tres aproximaciones de marcaje isotópico (SILAC, iTRAQ y ^{18}O) y diferentes espectrómetros de masas (LTQ, LTQ-Orbitrap y MALDI-TOF/TOF). Las muestras se repartieron entre cuatro grupos de investigación, ya que nuestro laboratorio no dispone de todos los espectrómetros de masas necesarios para este análisis. La preparación de cultivos (incluyendo el marcaje con SILAC) y la extracción de proteínas fueron realizados por el servicio de Proteómica de la Universidad Complutense de Madrid (en adelante, UCM). La digestión de los extractos utilizando la técnica de gel concentrante, el marcaje de los péptidos con ^{18}O e ITRAQ y su fraccionamiento por OffGel en 24 fracciones en un rango de pH de 3 a 10 fueron realizados en nuestro laboratorio (en adelante, CBMSO), desde donde se distribuyeron las muestras listas para su análisis por MS al resto de los laboratorios. Los análisis mediante LTQ de las muestra marcada con iTRAQ fueron realizados por el laboratorio de Proteómica del Instituto de Investigaciones Biomédicas de Barcelona (en adelante, IIBB). Los análisis con el equipo de alta resolución LTQ-Orbitrap fueron

realizados en colaboración con Juan Miguel Redondo y el servicio de Proteómica del Centro Nacional de Investigaciones Cardiovasculares (en adelante, CNIC). El análisis de las muestras marcadas con iTRAQ mediante MALDI-TOF/TOF fueron realizados en la UCM. Finalmente, los análisis de las muestras con el equipo LTQ y el procesamiento bioinformático posterior al análisis mediante espectrometría de masas fueron realizados íntegramente en el CBMSO. En la Figura R. 26 se muestra un esquema general de la estrategia experimental.

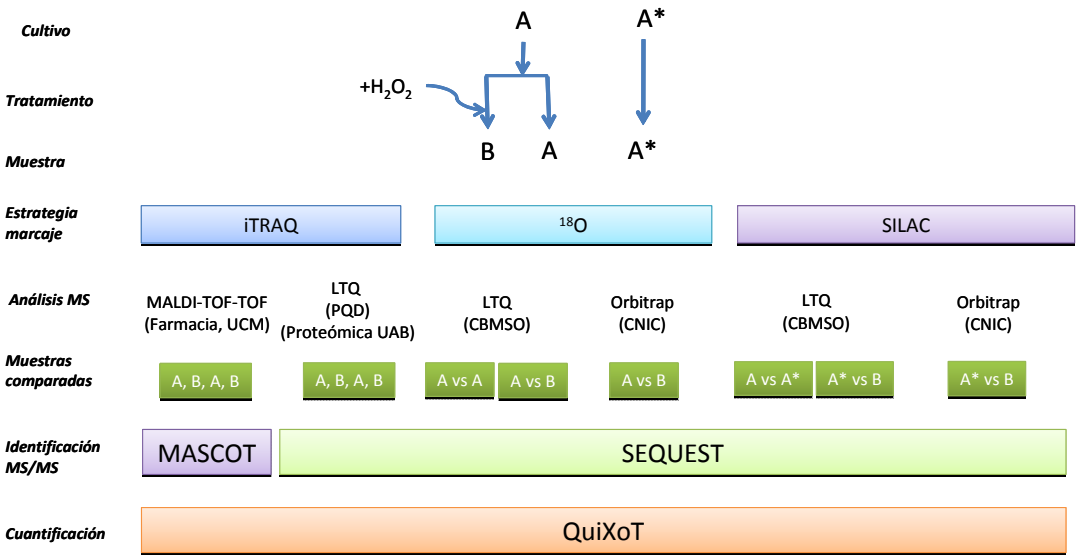


Figura R. 26 Estrategia experimental utilizada para el desarrollo de un modelo estadístico universal para marcaje con isótopos estables.

Los análisis de muestras marcadas con ^{18}O y SILAC en el equipo LTQ se realizaron de la misma manera que se describe en el capítulo anterior, es decir llevando a cabo un espectro de media resolución ("ZoomScan") en un rango estrecho de masas en torno a la masa del péptido precursor, que se utiliza para realizar la cuantificación relativa, seguido de un espectro de fragmentación MS/MS que se utiliza para identificar la secuencia del péptido. En iTRAQ se utilizaron los cuatro reactivos isobáricos (correspondientes a los iones "reporteros" (reporteros 114, 115, 116 y 117) de la siguiente manera: la muestra A se marcó con los reactivos que producen los iones reporteros 114 y 116, y la muestra B con los que producen los reporteros 115 y 117. En las muestras marcadas con iTRAQ y analizadas en el LTQ, los espectros de fragmentación se registraron en modo centroide, y las intensidades de los iones reporteros en cada espectro se utilizaron directamente como estimación de la concentración relativa de cada especie, sin otra transformación que la corrección isotópica descrita en

Material y Métodos. En las muestras marcadas con iTRAQ y analizadas por MALDI-TOF-TOF, se realizó un promediado de la intensidad de pico (*centroiding*) posterior a la adquisición de datos, que se procesaron entonces de la misma manera. En los análisis MS en el equipo LTQ-Orbitrap, las cuantificaciones se realizaron directamente en los espectros de barrido completo ("Full-scan"). El poder de resolución del equipo en este modo de barrido nos permitió utilizar como medida de intensidad de cada pico del espectro la intensidad más alta registrada en un estrecho rango (±0.01 Da) en torno a la razón masa/carga del ión esperado; este método minimiza el efecto de mezclas de especies co-eluidas.

Utilización de la información cromatográfica en alta resolución.

En las estrategias de alta resolución en las que se cuantifica sobre espectros MS de barrido completo es posible utilizar más de un espectro para la cuantificación, y no sólo aquél que corresponde al tiempo de retención donde se identifica el péptido. Por lo tanto, se puede aprovechar la información contenida en los espectros obtenidos a lo largo del pico cromatográfico completo de cada péptido analizado. Para detectar el ancho y el punto más intenso del pico cromatográfico correspondiente a cada péptido identificado se utilizó un programa (QuiXtoQuiX) realizado en el laboratorio por Marco Trevisan, y que es parte de su tesis doctoral. De esta manera, en cada pico cromatográfico se obtienen varios espectros cuantificables, y por lo general se obtiene más de un pico cromatográfico por cada péptido identificado (ya que la separación por OffGel no garantiza que cada fracción de IEF contenga todas las copias de un determinado péptido). Nos planteamos estudiar cuál es la mejor manera de integrar la información del pico, y cuánto podrían mejorar las cuantificaciones de las proteínas utilizando todo el pico con respecto a utilizar únicamente un espectro de cuantificación. En estudios preliminares (no mostrados) se observó que la varianza en la cuantificación de espectros correspondientes a un mismo pico cromatográfico es extremadamente baja (varianza intrapico $\sigma^2_S \approx 0.002$), siendo superior el error entre cuantificaciones del mismo péptido que se encontraran en picos cromatográficos diferentes (varianza interpico cromatográfico, $\sigma^2_C \approx 0.047$); por tanto el error cometido en las cuantificaciones de espectros puede descomponerse en dos: el error entre espectros correspondientes al mismos pico cromatográfico, y el error entre picos cromatográficos o fracciones diferentes donde se cuantifica el mismo péptido.

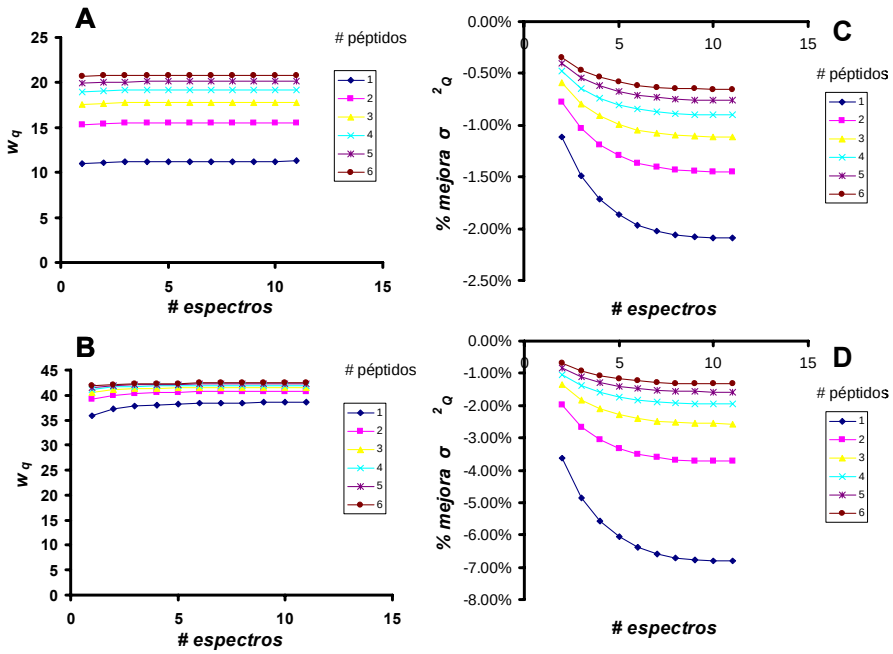


Figura R. 27 Efecto del número de medidas por pico cromatográfico sobre la varianza a nivel de proteína en un equipo de alta resolución. A partir de la varianza de la medida en el mismo pico cromatográfico ($\sigma^2_S \approx 0.002$), estimada previamente, se calcularon teóricamente las varianzas a nivel de proteína σ^2_Q que se obtendrían en un experimento de alta resolución de ^{18}O (A) y de SILAC (B), en función del número de espectros por pico cromatográfico y del número de péptidos por proteína, suponiendo que cada péptido aparece en dos picos cromatográficos diferentes por término medio. En C (^{18}O) y D

Para estimar cómo afecta el número de espectros utilizados por cada pico cromatográfico se realizó una simulación con datos preliminares de la estrategia ^{18}O analizada en Orbitrap. Con estos datos preliminares se estimaron las varianzas típicas intrapico e interpico ya mencionadas ($\sigma^2_S \approx 0.002$ y $\sigma^2_C \approx 0.047$), y se modeló mediante una función la pérdida de señal típica del segundo, tercer y cuarto espectros más intensos con respecto al espectro más intenso, de forma que pudiera estimarse el peso estadístico típico de un pico cromatográfico si se toman los *n* espectros más intensos. Con esta información, se simuló la varianza típica de las proteínas cuantificadas (usando las varianzas de péptido típicas de ^{18}O y SILAC) en función del número de espectros por pico cromatográfico, del número de picos cromatográficos por cada péptido, y del número de péptidos cuantificados por proteína (Figura R. 27). Los resultados observados indican que la utilización de más de un espectro por pico cromatográfico supone una mejora ínfima en la varianza de proteína, siempre inferior al 4%. Puesto que la utilización de más de un espectro por pico cromatográfico no influye de forma

decisiva en la calidad de la cuantificación, decidimos utilizar en los análisis definitivos únicamente el espectro más intenso de cada pico cromatográfico, por simplificación del modelo y mayor velocidad en el procesamiento de los datos.

Corrección de la conversión de arginina a prolina en marcaje SILAC

La tecnología SILAC no está del todo exenta de artefactos. Una cierta proporción de la arginina marcada isotópicamente (arginina pesada) puede ser transformada por un proceso metabólico en prolina pesada(SC Bendall et al., 2008) de forma que la señal de los péptidos con *n* prolinas de la muestra marcada se divide en *n+1* señales: la señal principal (y esperada si no se produjera este fenómeno) en la que ninguna prolina de los péptidos proviene de la transformación de una arginina pesada, y las *n* señales correspondientes a las especies conteniendo prolinas pesadas, separadas todas ellas entre sí por una distancia equivalente al del marcado (6 Da, en este caso). Puesto que la señal principal del péptido marcado disminuye debido a esta conversión, este efecto introduce artefactos en la cuantificación diferencial de los péptidos que contienen prolina.

Para poder corregir el efecto de esta conversión, definimos un factor *g* que corresponde a la fracción de prolinas no marcadas que permanecen en la muestra y no han sido marcadas metabólicamente a partir de arginina pesada. Si r_{qp} es el número de prolinas que contiene el péptido *p* proveniente de la proteína *q*, la concentración de péptido pesado conteniendo prolina no marcada disminuirá en un factor de $g^{r_{qp}}$. Ello aumentará el logaritmo en base dos de la razón de concentraciones a nivel de péptido en un sumando constante dado por $r_{qp} \cdot \phi$, siendo $\phi = -\log_2(g)$. Por tanto, esta conversión metabólica puede describirse introduciendo en el modelo de efectos aleatorios un efecto fijo dependiente del número de prolinas. El error cometido a nivel de los péptidos conteniendo r_{qp} prolinas, correspondiente al parámetro β_{qp} de la ecuación general 0-11, se distribuiría normalmente con media $r_{qp} \cdot \phi$ y varianza σ^2_P, es decir $\beta_{qp} \sim \mathrm{N}(\mathrm{r_{qp}} \cdot \phi, \sigma_P^2)$.

El parámetro ϕ se puede estimar de forma no sesgada por el método de la máxima verosimilitud (cuyo desarrollo se explica en el anexo A), dando lugar a la siguiente estima

$$\phi = \frac{\sum_q \sum_p w_{qp} \left(x_q - x_{qp} \right) \cdot r_{qp}}{\sum_q \sum_p w_{qp} \cdot r_{qp}^2}$$

Ecuación R. 21

donde el significado de los parámetros es el mismo que el descrito anteriormente. La proporción de prolinas no marcadas metabólicamente se puede calcular finalmente usando la siguiente ecuación: $g = 2^{-\phi}$.

El cálculo de ϕ se lleva a cabo de forma iterativa. Primero se calcula el valor medio de la medida a nivel de proteína utilizando los péptidos que no contienen prolina. En este punto se establece además un umbral de FDR_q de forma que las proteínas sujetas a un cambio de expresión muy grande no contribuyan tampoco al cómputo. Posteriormente, esos valores se utilizan para hacer una estima de ϕ de acuerdo a la Ecuación R. 21, y las medidas a nivel de proteína se vuelven a calcular, esta vez teniendo en cuenta la conversión a prolina. El proceso se repite hasta que la variación relativa de ϕ es menor que un valor previamente establecido.

Rendimiento de identificaciones y cuantificaciones de los experimentos

En la Tabla R. 2 se muestra un resumen del número de identificaciones (espectros, péptidos y proteínas) y cuantificaciones válidas conseguidas en los experimentos. Como se observa en esta tabla, en todos los análisis se consiguieron identificar más de 1,000 proteínas, y en todos los casos se logran cuantificar correctamente más del 70% de ellas. El número de proteínas cuantificadas osciló en torno al millar y el número de péptidos cuantificados en la mayoría de los casos es superior a 4.000. El número de medidas es, en todos los casos, superior a 10.000. Estos números permiten un modelado estadístico muy preciso en todos los casos.

			Datos de identificación			Datos de cuantificación			
Marcaje	MS	comparación	# espectros (Ns)	# péptidos (Np)	# proteínas (Nq)	Ns	Np	Nq	# cambios
iTRAQ	LTQ (PQD)	115 - 117 (B vs B)	17991	4390	1178	17110	3946	795	2
		117 - 116 (B vs A)				16129	3991	1021	18
		116 - 115 (A vs B)				16327	4033	1040	19
	MALDI-TOF-TOF	115 - 117 (B vs B)	3149	2142	616	2793	1982	608	0
		116 - 117 (A vs B)				2720	1942	608	1
		116 - 115 (A vs B)				2882	2033	614	2
18O	LTQ	A vs B	17783	4445	1145	12640	3177	830	8
		A vs A	1120	4379	13665	10791	3644	922	1
	Orbitrap	A vs B	14632	8011	1840	11147	6400	1525	13
SILAC	LTQ	A* vs B	20258	4306	1053	17172	3360	778	24
		A vs A*	12807	2599	740	10353	2032	540	9
	Orbitrap	A* vs B	28591	11798	2360	15614	7022	1461	18

Tabla R. 2 Rendimiento de identificaciones y cuantificaciones de todos los experimentos realizados. Se detalla también el número de cambios de expresión estadísticamente significativos detectado.

Pesos estadísticos de ajuste de cada aproximación

En todos los casos se observó que las medidas individuales, consideradas conjuntamente, no se pueden explicar utilizando una distribución normal, observándose desviaciones muy claras (no se muestra). Se intentó entonces, de la misma manera que se realizó en el apartado anterior, implementar un peso estadístico de ajuste que fuera capaz de estratificar las medidas de acuerdo a su varianza respectiva, de manera que se verificara la normalidad local en poblaciones de medidas con el mismo peso estadístico. Los pesos estadísticos de ajuste que finalmente se implementaron en cada uno de los casos se resumen en la Tabla R. 3.

Marcaje	Resolución	v_{qps}	Descripción de las sumas de cuadrados utilizadas
^{18}O	Baja	$\frac{I_{max}^2}{SQ_{PEP}+SQ_{\Delta max}}$	SQ_{PEP}: suma de cuadrados media en la envoltura isotópica del péptido. $SQ_{\Delta max}$: suma de cuadrados de una región de 2 Th en un lateral de la envoltura isotópica que depende la dirección del cambio de expresión.
	Alta	$\frac{I_{max}^2}{SQ}$	SQ: suma de cuadrados media de la envoltura isotópica.
SILAC	Baja	$\frac{I_{max}^2}{SQ_1 + SQ_2}$	SQ_1: suma de cuadrados media de la envoltura isotópica de la especie ligera del péptido. SQ_2: suma de cuadrados media de la envoltura isotópica de la especie pesada del péptido.
	Alta	$\frac{I_{max}^2}{SQ}$	SQ: suma de cuadrados media de la envoltura isotópica.
iTRAQ	Alta/Baja	I_{max}^2	

Tabla R. 3 Resumen de pesos estadisticos utilizados según la estrategia experimental. En todos los casos I_{max} representa la intensidad de la especie más abundante.

Tests de normalidad

En cada uno de los análisis realizados se comprobó que las medidas x_{qps}, una vez ordenadas por su peso estadístico v_{qps}, siguen una distribución normal de forma local. Para ello se utilizaron los tests de normalidad de D'Agostino(RB D'Agostino, 1971) ya utilizados en el capítulo anterior. En la figura Figura R. 28 se representan algunos resultados característicos de estos análisis de normalidad.

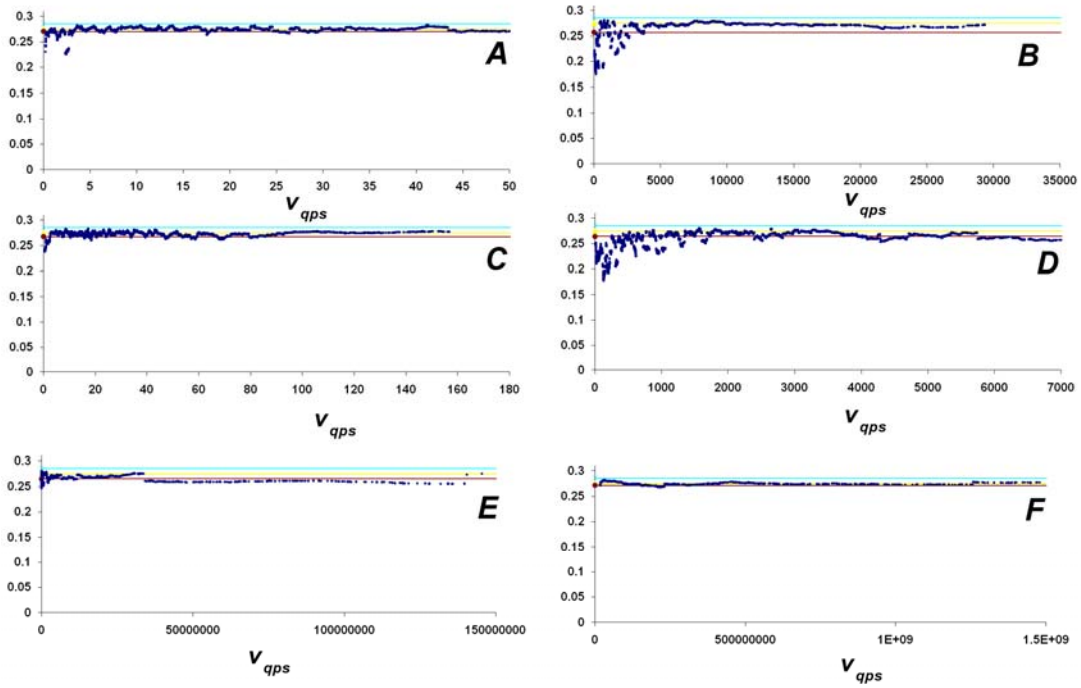


Figura R. 28 Tests de normalidad de D'Agostino. (A) ^{18}O + LTQ, (B) ^{18}O + Orbitrap, (C) SILAC + LTQ, (D) SILAC + Orbitrap, (E) iTRAQ + LTQ(PQD), (F) iTRAQ + MALDI-TOF-TOF

Cálculo de la constante k y de las varianzas de los tres niveles

En la Tabla R. 4 se detallan los valores de *k*, calculados para todos los experimentos mediante el mismo procedimiento que en el capítulo anterior. La constante *k* depende del método de cuantificación escogido, del peso estadístico y del aparato de medida, por lo que sus valores cambian de una aproximación a otra, pero adquiere valores consistentemente semejantes cuando se utiliza la misma aproximación en experimentos independientes.

Marcaje	MS	comparación	k calculado
iTRAQ	LTQ (PQD)	115 - 117 (B vs B)	700
		117 - 116 (B vs A)	700
		116 - 115 (A vs B)	700
	MALDI-TOF-TOF	115 - 117 (B vs B)	2×10^6
		116 - 117 (A vs B)	2×10^6
		116 - 115 (A vs B)	2×10^6
18O	LTQ	A vs B	0.17
		A vs A	0.17
	Orbitrap	A vs B	50
SILAC	LTQ	A* vs B	0.13
		A vs A*	0.10
	Orbitrap	A* vs B	16

Tabla R. 4 Valores de *k* calculados para todas las aproximaciones.

La estimación de las varianzas generales del experimento (σ^2_S, σ^2_P, σ^2_Q) se realizó de un modo diferente al descrito en el capítulo anterior. Debido a la diferente naturaleza de las muestras y al elevado número de datos acumulados, fue necesario desarrollar un método más robusto, basado en el cálculo de la mediana de las desviaciones absolutas y que corrige el efecto de sesgo en el cálculo de las varianzas utilizando una corrección local de grados de libertad. Este método, que es mucho más robusto que el método anterior y no se ve apenas afectado por las desviaciones atípicas de la muestra, se describe en el anexo A. Las varianzas calculadas para cada experimento se muestran en la Tabla R. 5, y los límites de confianza han sido calculados mediante la simulación explicada en el capítulo anterior.

Marcaje	MS	comparación	Varianza de espectros (σ^2_S) (95% C.I.)	Varianza de péptidos (σ^2_P) (95% C.I.)	Varianza de proteínas (σ^2_Q) (95% C.I.)	corrección Arg -> Pro en SILAC (ϕ)
iTRAQ	LTQ (PQD)	115 - 117 (B vs B)	0,103 (0,001 - 0,2)	0,036 (0,001 - 0,2)	0,01 (0 - 0,01)	
		117 - 116 (B vs A)	0,141 (0,001 - 0,195)	0,106 (0,001 - 0,2)	0,01 (0 - 0,01)	
		116 - 115 (A vs B)	0,125 (0,001 - 0,2)	0,053 (0,001 - 0,2)	0,01 (0 - 0,01)	
	MALDI-TOF-TOF	115 - 117 (B vs B)	0,009 (0,001 - 0,2)	0,126 (0,001 - 0,2)	0,01 (0 - 0,01)	
		116 - 117 (A vs B)	0,017 (0,001 - 0,2)	0,138 (0,001 - 0,2)	0.04 (0.03 - 0.05)	
		116 - 115 (A vs B)	0,003 (0,001 - 0,2)	0,008 (0,001 - 0,2)	0,01 (0 - 0,01)	
18O	LTQ	A vs B	0,003 (0,001 - 0,004)	0,022 (0,016 - 0,031)	0,01 (0 - 0,01)	
		A vs A	0,019 (0,014 - 0,022)	0,067 (0,047 - 0,092)	0,01 (0 - 0,01)	
	Orbitrap	A vs B	0,008 (0,001 - 0,015)	0,044 (0,001 - 0,162)	0.04 (0.03 - 0.05)	
SILAC	LTQ	A* vs B	0,004 (0,002 - 0,005)	0,007 (0,004 - 0,012)	0.02 (0.01 - 0.03)	0.025
		A vs A*	0,006 (0,002 - 0,005)	0,008 (0,004 - 0,013)	0.01 (0 - 0.02)	0.030
	Orbitrap	A* vs B	0 (0,001 - 0,007)	0,002 (0,001 - 0,023)	0.02 (0 - 0.02)	0.036

Tabla R. 5 Estimación de varianzas σ^2_S (A), σ^2_P (B) y σ^2_Q (C). Se muestran las varianzas calculadas de los experimentos del proteoma de levadura reallizados para demostrar la generalización del modelo.

A nivel de medida, la varianza σ^2_S más alta corresponde a la aproximación de iTRAQ analizada en LTQ (en modo PQD), lo que se encuentra dentro de lo previsto, dada la relativamente baja eficiencia de fragmentación del LTQ en la región de masas donde aparecen los iones reporteros. En el caso de iTRAQ analizado en un equipo de fragmentación cuadrupolar (MALDI-TOF-TOF), la varianza de la medida es mucho menor, ya que la eficiencia de fragmentación en este caso es mucho más alta y los iones reporteros están entre los más intensos del espectro de fragmentación, lo que se traduce en un error de cuantificación considerablemente menor. Estas varianzas están en el mismo orden que las obtenidas para la cuantificación en modo MS en un equipo de baja resolución (SILAC y ^{18}O en LTQ). Consistentemente, la menor varianza se consigue en la cuantificación en modo MS en un equipo de alta resolución (SILAC y ^{18}O en Orbitrap).

Las dispersiones a nivel de péptido (varianzas σ^2_P) son generalmente superiores en el método iTRAQ, que para otras aproximaciones de marcaje, si bien se observa una gran variabilidad en las varianzas al hacer la comparativa entre diferentes parejas de reporteros, mientras que las varianzas obtenidas en el marcaje con ^{18}O son mucho más parecidas entre sí. Puesto que la etapa de digestión sigue el mismo protocolo en todos los casos, esas diferencias sugieren que el marcaje con el reactivo iTRAQ no se ha alcanzado de forma homogénea en todas las muestras, aumentando la varianza a nivel de péptido. En el caso de la aproximación de SILAC, las varianzas de péptido calculadas son en todos los casos mínimas, muy próximas a cero, como corresponde a un método en el que el marcaje se produce de forma previa a la digestión.

Finalmente, las dispersiones a nivel de proteína (varianzas σ^2_Q) son esencialmente semejantes en todos los casos, de forma independiente de la aproximación utilizada y prácticamente en ningún caso se hacen significativamente diferentes de cero. No se detectan diferencias apreciables entre los casos donde se compara el mismo extracto (B vs B, A vs A), cuya varianza debería ser teóricamente nula, y los casos donde se comparan muestras diferentes (A vs B, A vs A*), sugiriendo que la manipulación de la muestra (cultivo y extracción de proteínas) no introduce fuentes de error apreciables

Análisis de valores atípicos de cuantificaciones de espectros y péptidos

Como paso previo a la detección de cambios de expresión, se eliminaron aquellos valores de cuantificación atípicos en cada uno de los niveles estudiados. Para ello se clasificaron, como se hizo en el capítulo anterior, aquellas cuantificaciones con un valor de FDR bajo (menor al 5%) a nivel de espectro y de péptido. En este proceso se diferenciaron aquellas cuantificaciones en las que se pudo observar algún problema en la cuantificación, como un co-eluído que interfiere en la envoltura isotópica del péptido cuantificado o un mal ajuste de la envoltura teórica, de aquéllas en las que no se encuentra ninguna razón que justifique un valor desviado de la media y que se consideran realmente atípicas. Es importante tener en cuenta que no debería encontrarse un número muy grande de cuantificaciones realmente atípicas, ya que en caso contrario se estaría desaprovechando un volumen importante de la información del experimento, y podría ser signo de un planteamiento erróneo del modelo estadístico de hipótesis nula. Además, un número suficientemente bajo de cuantificaciones atípicas no justificadas permitiría automatizar el modelo mediante un algoritmo iterativo de eliminación de cuantificaciones por debajo de un cierto valor de FDR, ya que el volumen de información perdida sería mínimo.

Valores atípicos de cuantificaciones a nivel de espectro

En la Tabla R. 6 se detalla el número y porcentaje con respecto al total de cuantificaciones atípicas a nivel de medida (FDR_{qps} inferior al 5%). En los casos de iTRAQ, no es posible justificar las cuantificaciones atípicas, ya que la medición se realiza en espectros de fragmentación MS^2 mediante iones reporteros. En todos los demás casos, el número de espectros atípicos no justificados es siempre igual o inferior al 0.5%, por lo que la pérdida de información al eliminar estos espectros del estudio es mínima. Para mostrar la robustez del método, en la Figura R. 29 se representan, para varios experimentos representativos, la desviación de cada espectro con respecto a su media (x_{qps}-x_{qp}) tras eliminar las cuantificaciones atípicas justificables, y las gráficas de normalidad que muestran que tras la eliminación de todas las cuantificaciones atípicas, las muestras siguen una distribución normal. En todos los casos, incluso aquéllos que no se muestran, se observaron distribuciones plenamente compatibles con una curva normal.

Marcaje	MS	comparación	Ns	# cuantificaciones de espectro atípicas	% cuantificaciones de espectro atípicas
iTRAQ	LTQ (PQD)	115 - 117 (B vs B)	17110	109	0,64%
		117 - 116 (B vs A)	16129	103	0,64%
		116 - 115 (A vs B)	16327	125	0,77%
	MALDI-TOF-TOF	115 - 117 (B vs B)	2793	9	0,32%
		116 - 117 (A vs B)	2720	3	0,11%
		116 - 115 (A vs B)	2882	11	0,38%
18O	LTQ	A vs B	12640	71	0,56%
		A vs A	10791	31	0,29%
	Orbitrap	A vs B	11147	22	0,20%
SILAC	LTQ	A* vs B	14143	49	0,35%
		A vs A*	10353	13	0,13%
	Orbitrap	A* vs B	15614	13	0,08%

Tabla R. 6 Número y porcentaje de cuantificaciones atípicas a nivel de medida. Se cuentan como cuantificaciones atípicas aquellas que tienen asociada una FDR_{qps} <5%.

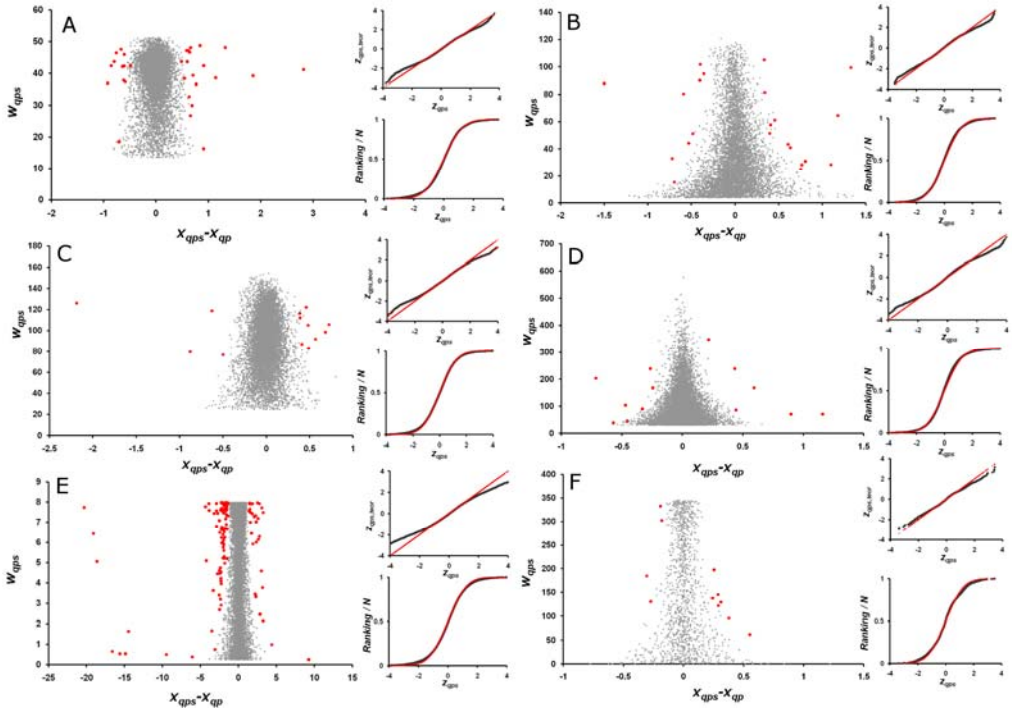


Figura R. 29 Detección de cuantificaciones atípicas a nivel de espectro. Se representa el peso estadístico de espectro w_{qps} frente a la dispersión (x_{qps}-x_{qp}), destacando las cuantificaciones atípicas (FDR_s<5%), y las correspondientes gráficas de normalidad. Se muestra un proteoma para cada estrategia experimental seguida (tipo de marcaje + MS). (A) 18O + LTQ, (B) 18O + Orbitrap, (C) SILAC + LTQ, (D) SILAC + Orbitrap, (E) iTRAQ + LTQ(PQD), (F) iTRAQ + MALDI-TOF-TOF.

Valores atípicos de cuantificaciones a nivel de péptido

De manera análoga, se estudiaron los valores atípicos de cuantificaciones de péptidos con respecto a sus medias (cuantificaciones de proteína). Para este estudio, se utilizó un protocolo de procesado de muestras para marcaje isotópico mejorado (E Bonzon-Kulichenko et al., en revisión) respecto al método usado en el apartado anterior, en el que las digestiones parciales y las metioninas oxidadas no afectan a las cuantificaciones (datos no mostrados). El número de cuantificaciones atípicas de péptido no justificables es mínimo en todos los casos (menor al 0.6%) como se muestra en la Tabla R. 7. En la **Error! Reference source not found.** se representaron los mismos experimentos de la Figura R. 29, mostrando las desviaciones de cada péptido con respecto a su media (x_{qp}-x_q) tras retirar las cuantificaciones atípicas justificables, y las gráficas de normalidad tras la eliminación de todas las cuantificaciones atípicas. Como puede observarse, todas las muestras representan correctamente a una distribución normal a nivel de péptido.

Marcaje	MS	comparación	Np	# cuantificaciones de péptido atípicas	% cuantificaciones de péptido atípicas
iTRAQ	LTQ (PQD)	115 - 117 (B vs B)	3946	20	0,51%
		117 - 116 (B vs A)	3991	4	0,10%
		116 - 115 (A vs B)	4033	6	0,15%
	MALDI-TOF-TOF	115 - 117 (B vs B)	1982	0	0,00%
		116 - 117 (A vs B)	1942	0	0,00%
		116 - 115 (A vs B)	2033	2	0,10%
18O	LTQ	A vs B	3177	2	0,06%
		A vs A	3644	6	0,16%
	Orbitrap	A vs B	6400	25	0,39%
SILAC	LTQ	A* vs B	3360	14	0,42%
		A vs A*	2032	7	0,34%
	Orbitrap	A* vs B	7022	20	0,28%

Tabla R. 7 Número y porcentaje de cuantificaciones atípicas a nivel de p'eptido**. Se cuentan como cuantificaciones atípicas aquellas que tengan asociada una FDR_P <5%.**

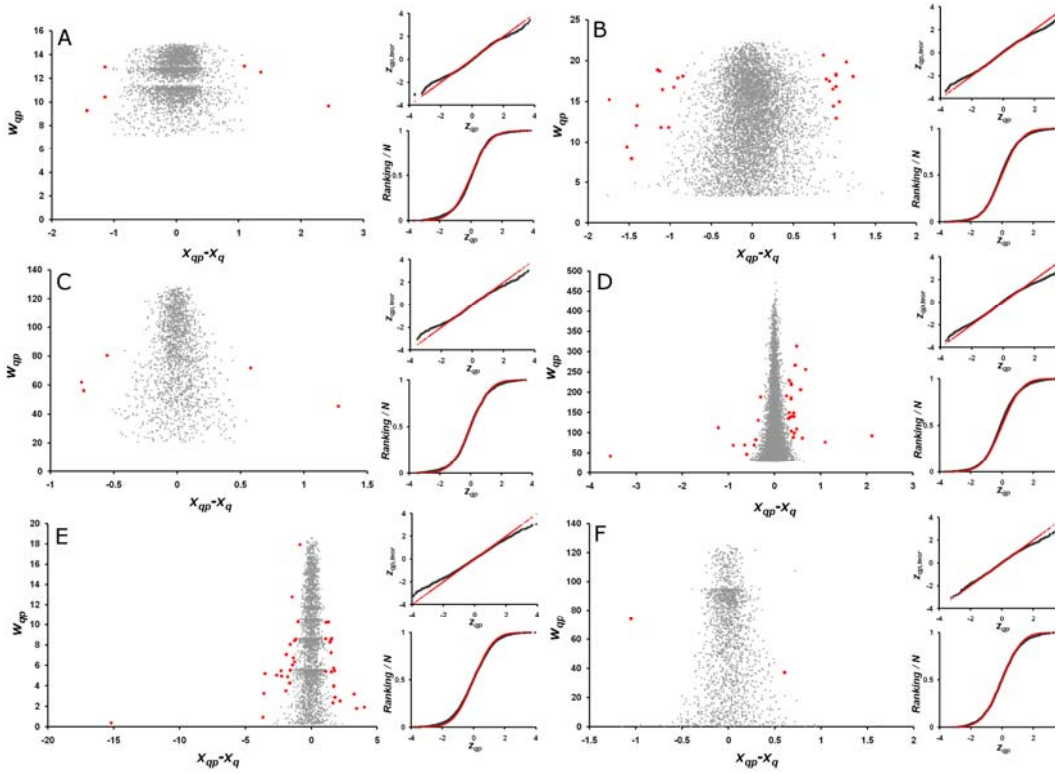


Figura R. 30 Detección de cuantificaciones atípicas a nivel de péptido. Se representa el peso estadístico de péptido w_{qp} frente a la dispersión (x_{qp}-x_q), destacando las cuantificaciones atípicas de péptido (FDR_{qp}<5%), y las correspondientes gráficas de normalidad. Se muestra un proteoma para cada estrategia experimental seguida (tipo de marcaje + MS). (A) ^{18}O + LTQ, (B) ^{18}O + Orbitrap, (C) SILAC + LTQ, (D) SILAC + Orbitrap, (E) iTRAQ + LTQ(PQD), (F) iTRAQ + MALDI-TOF-TOF.

Análisis de hipótesis nulas y cambios de expresión

Se comprobaron los cambios de expresión encontrados en cada uno de los experimentos, así como la normalidad de los datos a nivel de proteína. En el caso de las hipótesis nulas, se observa que no hay apenas cambios de expresión falsos, excepto en el caso de la aproximación de SILAC. Estos cambios de expresión son probablemente debidos a que la hipótesis nula de SILAC debe realizarse con dos cultivos diferentes (muestras A y A*), y mínimas diferencias en el crecimiento de ambos cultivos pueden causar diferencias de expresión grandes en algunas proteínas. La Figura R. 31 muestra todos los cambios de expresión detectados en todos los experimentos, además de las correspondientes gráficas de normalidad que manifiestan que todos los experimentos cumplen con la hipótesis de normalidad exigida a nivel de proteína. Los cambios observados son coherentes con el tipo de tratamiento realizado en la muestra B, y existe una tendencia muy clara a repetirse las mismas proteínas en las diferentes estrategias (marcaje isotópico + espectrómetro de masas), aunque lógicamente en unos casos la profundidad obtenida es superior a la de los otros. Un resumen completo de los cambios de expresión encontrados con un FDRq < 5% se detalla en el anexo A.

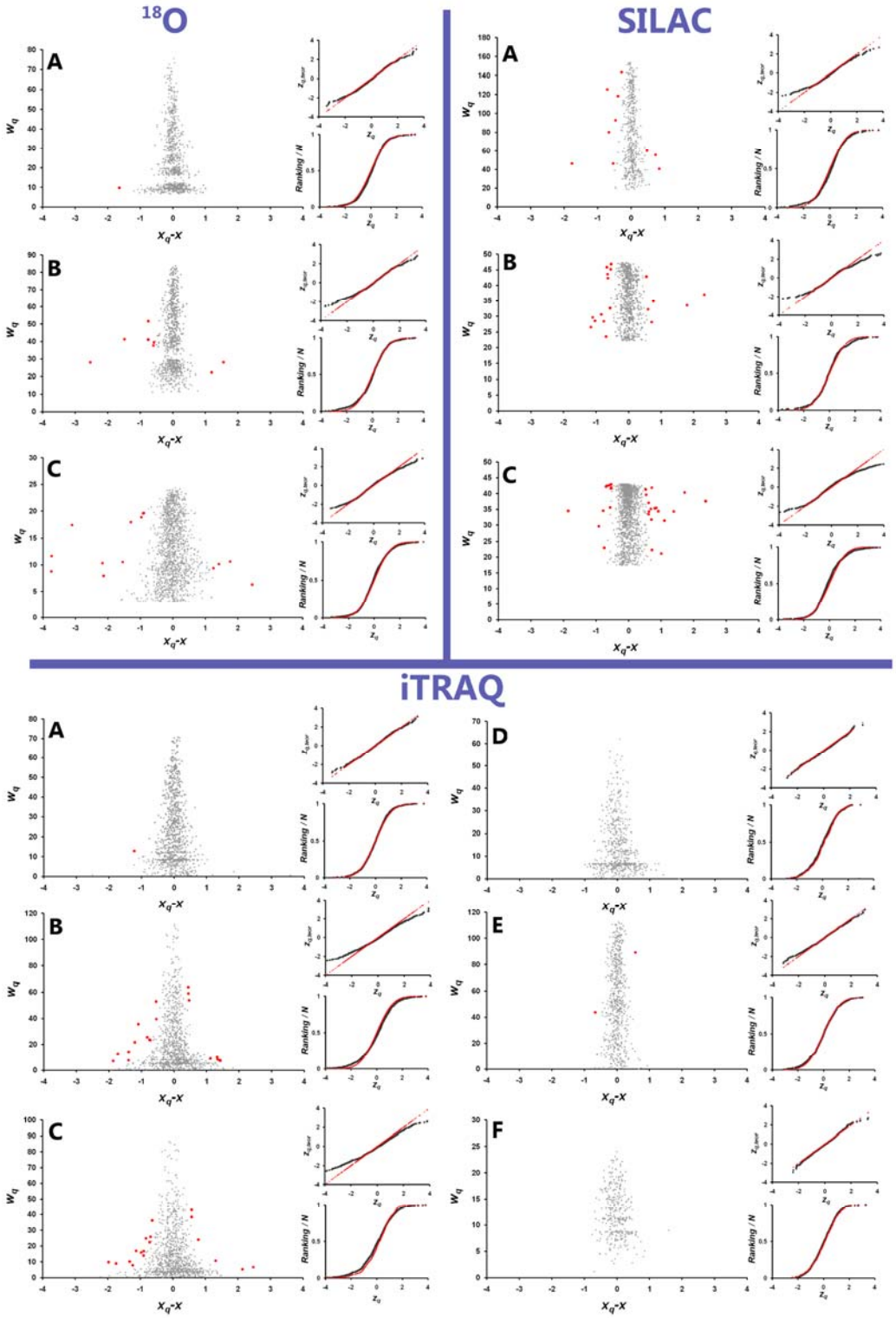


Figura R. 31 Detección de cambios de expresión. Se destacan los cambios de expresión (FDR_q < 5%) de todos los proteomas analizados. (muestras comparadas + MS). ^{18}O: (A) A vs A + LTQ, (B) A vs B + LTQ, (C) A vs B + Orbitrap. SILAC: (A) A vs A* + LTQ, (B) B vs A* + LTQ, (C) B vs A* + Orbitrap. iTRAQ: (A) B vs B (117 vs 115) + LTQ(PQD), (B) A vs B (116 vs 115) + LTQ(PQD), (C) A vs B (116 vs 117) + LTQ(PQD), (D) B vs B (117 vs 115) + MALDI-TOF-TOF, (E) A vs B (116 vs 115) + MALDI-TOF-TOF, (F) A vs B (116 vs 117) + MALDI-TOF-TOF

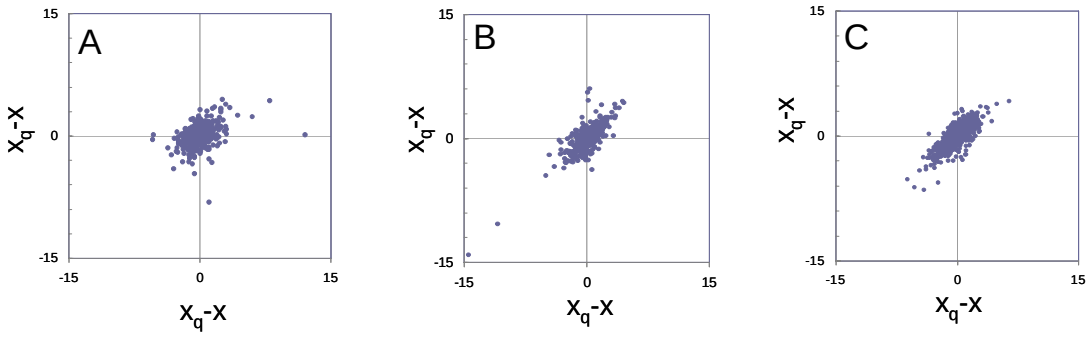


Figura R. 32 Comparativa entre los resultados de cuantificación diferencial**. Se muestra una comparación de los resultados de cuantificación diferencial a nivel de proteína en experimentos obtenidos mediante diferentes marcajes isotópicos y espectrómetros de masas: (A) SILAC + LTQ (hipótesis nula) vs ^{18}O + LTQ (hipótesis nula). (B) SILAC + LTQ (B vs A*) vs SILAC + Orbitrap (B vs A*), (C) iTRAQ + LTQ (PQD) (116 vs 115) (A vs B) vs iTRAQ + LTQ (PQD) (117 vs 116) (A vs B)**

Al comparar los diversos experimentos entre sí, se observa una coherencia entre los cambios de expresión observados, como puede verse en la Figura R. 32, donde se han representado algunas de estas comparativas a modo de ejemplo. En la comparación de experimentos de hipótesis nula se observa un resultado equiprobable para las cuantificaciones de las proteínas comunes, no existiendo correlación apreciable entre las medidas (veáse Figura R. 32 A). En cambio, en la comparativa de experimentos en los que se sometió una de las muestras al tratamiento con agua oxigenada, se observa una clara correlación entre las medidas de cuantificación a nivel de proteína (Figura R. 32 B y C).

3. Integración de los algoritmos en una plataforma de software de proteómica cuantitativa (QuiXoT).

El continuo desarrollo de herramientas y algoritmos matemáticos destinados a proteómica cuantitativa hace necesario el desarrollo de una plataforma de software que sirva tanto de banco de pruebas de los algoritmos propuestos, como de instrumento de análisis final de los diferentes experimentos realizados en el laboratorio. En este apartado se describe el diseño de la plataforma de software para proteómica cuantitativa desarrollada en esta tesis, que contiene todos los algoritmos estadísticos descritos, y que hemos denominado QuiXoT.

3.1 Flujo de datos en proteómica cuantitativa

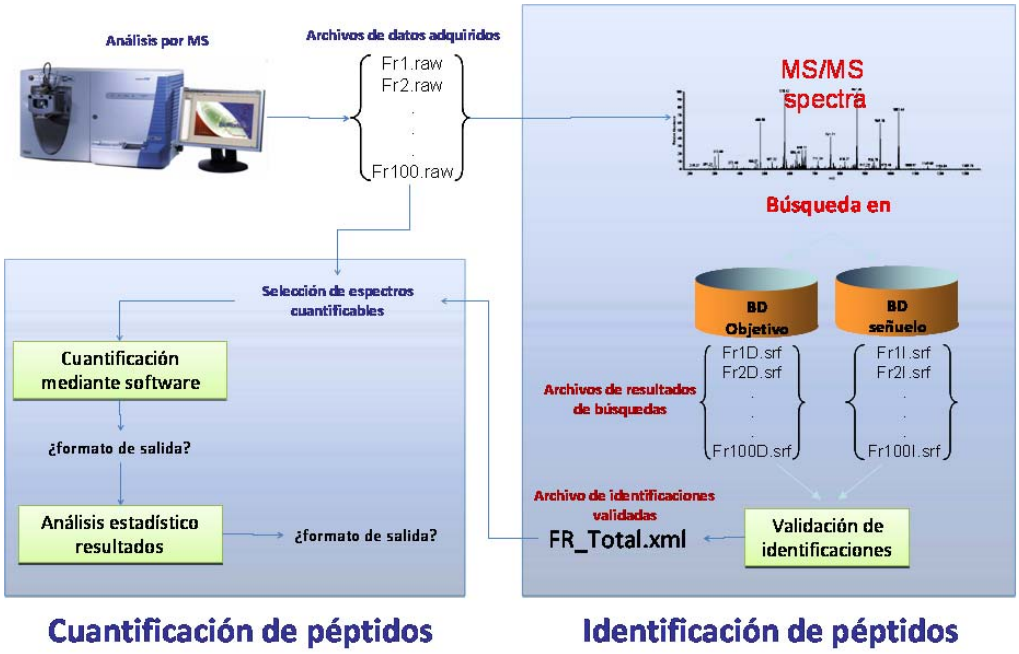


Figura R. 33 Flujo de datos típico en un experimento de proteómica cuantitativa.

La plataforma se diseñó para ser capaz de analizar cualquier experimento de proteómica cuantitativa mediante marcaje isotópico estable. En este tipo de experimentos se utiliza siempre un esquema de flujo de datos similar al mostrado en la Figura R. 33. Los datos obtenidos mediante espectrometría de masas son adquiridos en alguno de los formatos proporcionados por el fabricante del equipo. De ellos se debe de obtener la información de identificación de cada péptido (o proteína) que se desee cuantificar, y extraer la colección de espectros que va a ser utilizada para la cuantificación. Todo esto es utilizado por el software que realiza la cuantificación de los pares de péptidos o proteínas del experimento, y después

debe realizarse un tratamiento estadístico para discernir qué cambios de expresión son estadísticamente significativos.

Para almacenar los datos de identificación – tasa de error de identificación, puntuaciones asignadas por el o los motores de búsqueda utilizados, y otros datos relacionados con la secuencia –, los datos crudos de cuantificación – intensidades relativas detectadas, bondades de ajuste… – y los datos del análisis estadístico realizado, se desarrolló un esquema de XML, llamado QuiXML, cuyos requisitos y características se describen en el anexo A.

3.2 Desarrollo de un módulo de cuantificación para marcaje isotópico con ^{18}O en trampa lineal

El primer módulo de QuiXoT que se desarrolló en esta tesis es capaz de leer archivos de identificación QuiXML, y extraer los espectros ZoomScan de los archivos .RAW de los equipos de Thermo, que contienen todos los espectros de masas adquiridos por el equipo a lo largo de cada carrera cromatográfica. Este módulo cuantifica los espectros de péptidos marcados con ^{18}O siguiendo un algoritmo de ajuste iterativo de Newton-Gauss sobre una función, previamente desarrollada en el laboratorio(A Ramos-Fernandez et al., 2007), que permite el cálculo de la eficiencia de marcaje para cada espectro cuantificado, a la vez que determina los valores de cuantificación relativa de cada péptido, tal y como se explica en material y métodos.

Los modelos estadísticos analizados en esta tesis doctoral fueron incorporados en QuiXoT, de forma que todo el proceso de análisis de datos de un experimento de ^{18}O puede llevarse a cabo con este software. El programa realiza representaciones gráficas automáticas de cualquier parámetro del archivo QuiXML con control dinámico sobre los datos mediante filtros y funciones de zoom, lo que permite la inspección de los elementos que se consideren clave para el análisis, como por ejemplo la eficiencia de marcaje de un experimento de ^{18}O, el grado de digestión parcial o la presencia de queratinas.

3.3 Generalización de QuiXoT para el análisis de experimentos obtenidos por cualquier método de marcaje isotópico

El estudio de otros tipos de marcaje isotópico (iTRAQ, SILAC…) y otros espectrómetros de masas y el desarrollo de un modelo estadístico general con procedimientos específicos para cada una de las aproximaciones hizo necesaria una ampliación del software original de cuantificación QuiXoT, convirtiéndolo en una plataforma general de software de proteómica cuantitativa, descrita en la Figura R. 34.

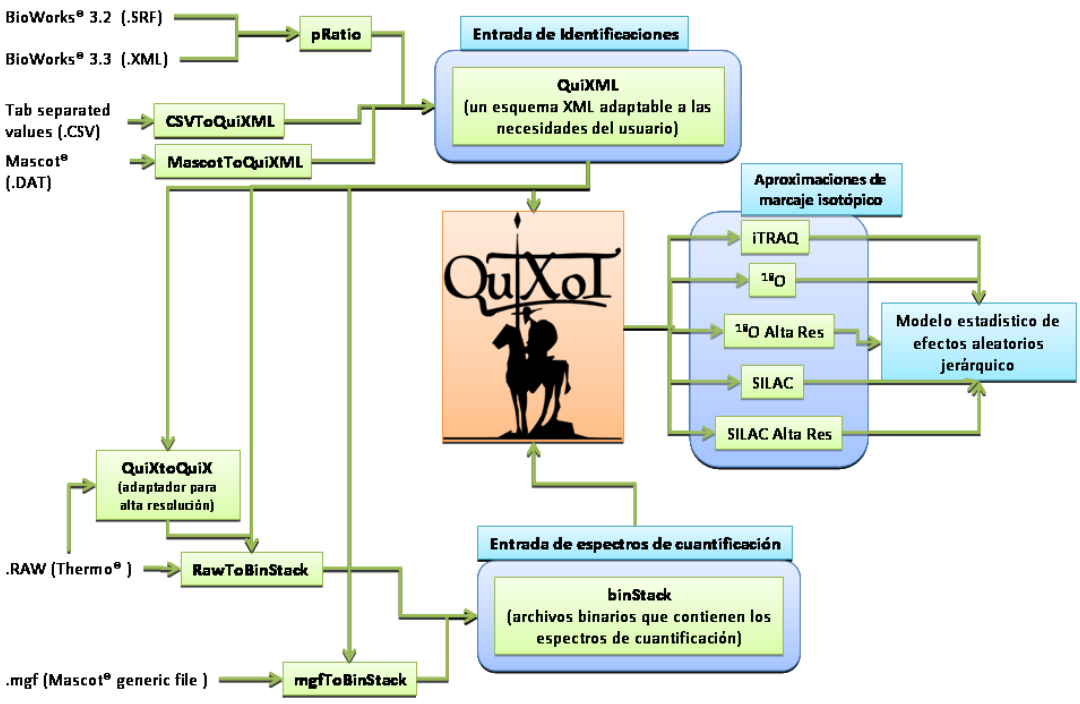


Figura R. 34 Esquema de la plataforma de software QuiXoT.

Se adaptó un formato de plataforma capaz de adquirir datos de identificación de cualquier motor de búsqueda, utilizando para ello tres herramientas diferentes: pRatio, mascotToQuiXML y txtToQuiXML. El programa pRatio valida identificaciones de SEQUEST mediante el método de la razón de probabilidades, incluyendo el punto isoeléctrico como parámetro estadístico, y utiliza el método refinado de cálculo de tasas de error. El programa mascotToQuiXML convierte identificaciones realizadas por el motor de búsqueda Mascot a un archivo en el formato QuiXML. El programa txtToQuiXML transforma – mapea – cualquier tipo de tabla de datos de identificaciones en un archivo QuiXML, por lo que es válido para cualquier otro motor de búsqueda o método de identificación que se utilice, proporcionando la

información fundamental de la identificación (identificación de la proteína, secuencia identificada, y espectro o espectros de identificación).

La adquisición de los espectros de cuantificación puede realizarse mediante dos programas: rawToBinStack y mgfToBinStack. Ambos extraen los espectros de cuantificación requeridos por una lista (un archivo QuiXML conteniendo los datos de identificación) y los convierten al formato binario utilizado por QuiXoT (llamado binStack). RawToBinStack extrae espectros de archivos de carreras .RAW de equipos de Thermo, y mgfToBinStack extrae los espectros de archivos en Mascot Generic File (.mgf), un formato ampliamente utilizado en proteómica. Además de estos métodos de extracción de espectros, se implementó en la plataforma un programa adaptador para alta resolución, llamado QuiXtoQuiX y programado por Marco Trevisan en el laboratorio, que busca en los archivos .RAW de Thermo todos los espectros correspondientes al pico cromatográfico de cada una de las secuencias identificadas de un archivo QuiXML.

QuiXoT fue adaptado para analizar los datos de cuantificación sobre varias aproximaciones de marcaje isotópico: iTRAQ, SILAC y ^{18}O. La aproximación de iTRAQ puede ser utilizada también para marcaje mediante TMT, y la aproximación de SILAC puede ser configurada para poder utilizarse con cualquier tipo de marcaje isotópico estable basada en el cálculo de la diferencia relativa de intensidades de una especie ligera y una pesada medidas en un mismo espectro MS. El método de ^{18}O se mantuvo de forma independiente, ya que requiere un algoritmo particular que calcula la eficiencia de marcaje para cada espectro cuantificado, como se explicó anteriormente. Además, se implementaron de forma independiente dos métodos para alta resolución (^{18}O alta resolución y SILAC alta resolución), ya que el cálculo de las intensidades correspondientes a las especies observada en los espectros se lleva a cabo de forma diferente. La Figura R. 35 muestra un ejemplo de cuantificación de cada una de estas aproximaciones. El método de validación estadística es común para todas las aproximaciones, con la ligera excepción en los métodos de SILAC, que incluyen en el modelo una función de corrección para los péptidos con prolina, como se ha descrito anteriormente.

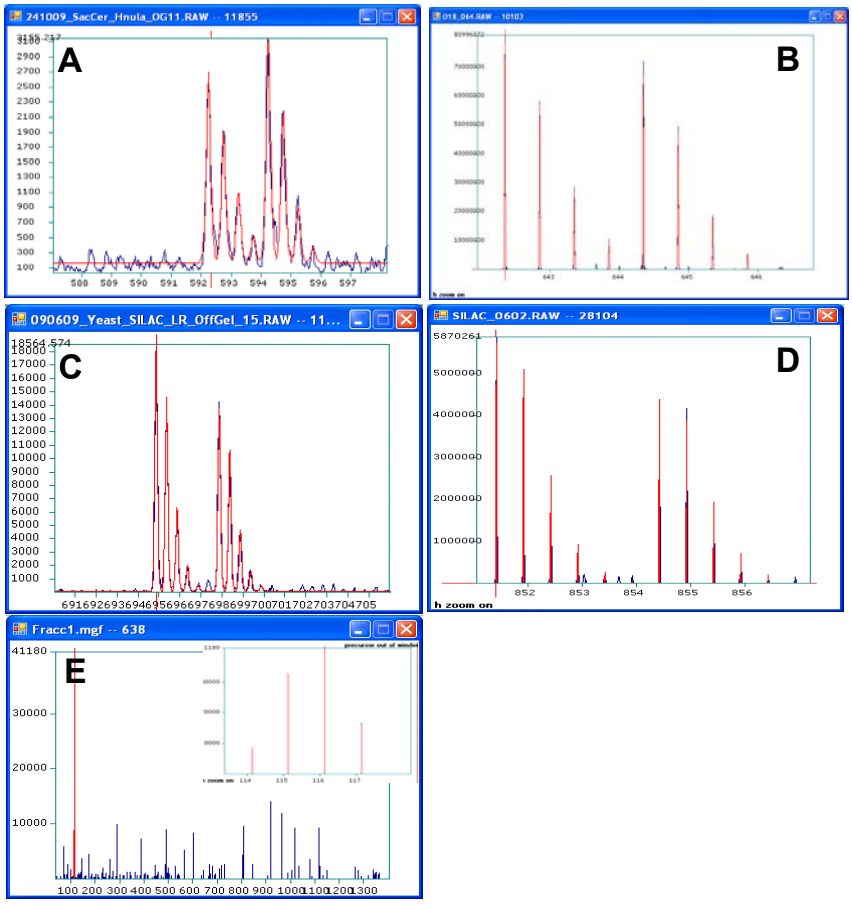


Figura R. 35 Ejemplos de cuantificaciones analizadas mediante QuiXoT. (A) ^{18}O en baja resolución, (B) ^{18}O en alta resolución, (C) SILAC en baja resolución, (D) SILAC en alta resolución, y (E) iTRAQ.

El programa permite configurar por el usuario prácticamente todos los aspectos relativos a la estructura química de los péptidos (composición atómica de los aminoácidos y de las modificaciones usadas), la cuantificación (tolerancias de intensidad, ciclos de iteración utilizados en rutinas de Newton-Gauss, etc.) y los relativos al estudio estadístico a abordar (varianzas, filtros sobre los datos a analizar, tolerancias de tasas de error…), lo que hace posible tanto un análisis profundo y controlado (con el que poder hacer pruebas sobre el modelo estadístico aplicado) como un análisis rápido y automático, destinado a los usuarios finales del software.

Discusión

Avances en algoritmos de identificación masiva de péptidos

El método de la Razón de Probabilidades

Prácticamente todos los métodos descritos en la bibliografía para realizar inferencia estadística en la identificación de péptidos usando SEQUEST están basados en el análisis de un conjunto de puntuaciones (típicamente las puntuaciones Xcorr) obtenidas al buscar en bases de datos de secuencias un largo conjunto de espectros MS/MS (A Keller et al., 2002, T Kislinger et al., 2003, D Lopez-Ferrer et al., 2004, MJ MacCoss et al., 2002, RE Moore et al., 2002). Explicado de forma concisa, las mejores puntuaciones, sean tomadas como sean, son agrupadas, ordenadas de mejor a peor – un *ranking* – y utilizadas de esta forma para construir una distribución acumulativa de puntuaciones, que es utilizada para determinar la significatividad estadística de la identificación peptídica. Repitiendo el proceso con una base de datos invertida, estas distribuciones permiten calcular a su vez el error cometido en el grupo de péptidos identificados en términos de FDR. Estas distribuciones acumulativas son estimaciones estadísticas de la distribución promedio de probabilidad (*average-score*) de la mejor puntuación, una función que evalúa la probabilidad de encontrar un espectro con puntuación igual o mejor que la puntuación observada.

Otro punto de vista estadístico diferente para la identificación de péptidos consiste en evaluar la significatividad individual de cada una de las identificaciones, de forma aislada e independiente de las demás, evaluando la probabilidad de que la puntuación obtenida sea igual o mejor que la mejor que se obtendría al comparar al azar el espectro con una colección de secuencias aleatorias del mismo tamaño. Esta probabilidad se evalúa construyendo las distribuciones individuales (o *single-spectrum*) de puntuaciones sobre el conjunto de secuencias contra las que se compara en la base de datos. Este concepto es semejante al del valor de expectación (o e-value), usado con frecuencia para buscar información en bases de datos (por ejemplo mediante los algoritmos FASTA o BLAST) (SF Altschul et al., 1990, D Fenyo,RC Beavis, 2003, LY Geer et al., 2004, RG Sadygov,JR Yates, 3rd, 2003).

En el artículo publicado en el que se expone este trabajo(S Martinez-Bartolome et al., 2008) se analizan teóricamente las propiedades de las distribuciones individuales de los espectros, que, como se ha dicho, son funciones características de cada espectro analizado, expresándose matemáticamente la distribución promedio como una combinación de las distribuciones de los espectros individuales (esta parte de modelado teórico fue realizada por Fernando Martín-Maroto). La razón fundamental para expresar las distribuciones promedio en función de las distribuciones de los péptidos individuales subyacentes es que estas últimas pueden ser muy diferentes para diferentes espectros, por lo que no se puede esperar que grandes conjuntos de datos tengan un comportamiento estadístico homogéneo.

Una propiedad importante validada mediante simulación en dicho trabajo teórico es que el valor de la distribución media para una cierta puntuación refleja de manera fidedigna la fracción de espectros del conjunto que tienen una calidad mejor, es decir, que tienden a producir una puntuación mejor simplemente por azar. Este término de "calidad" ha sido aplicado previamente a espectros MS/MS refiriéndose a las propiedades de los espectros que evalúan la probabilidad de que estos espectros se generen por la fragmentación de un péptido real(M Bern et al., 2004, EF Strittmatter et al., 2004). En este estudio sin embargo la calidad es introducida por Fernando Martín-Maroto(S Martinez-Bartolome et al., 2008) como un concepto matemático cuya interpretación práctica es que los espectros de mayor calidad son aquéllos que tienden a producir mayores puntuaciones cuando éstos son buscados en una base de datos de secuencias señuelo o aleatorias, y que permite extrapolar las propiedades de las distribuciones individuales a la distribución promedio. Esta extrapolación, que es tanto más válida cuanto mayor sea el tamaño de la colección de candidatos utilizada para puntuar un espectro, permite determinar la probabilidad de obtener la mejor puntuación cuando se ha obtenido una segunda mejor puntuación, que toma la forma de la razón de probabilidades comentada en la introducción. Este proceder integra de forma coherente el uso de los dos tipos de distribuciones, ya que la razón de probabilidades utiliza las distribuciones promedio de las dos mejores puntuaciones de cada espectro para realizar un ranking, y para cada espectro analizado se evalúa la razón entre la probabilidad de la primera puntuación de ser un PSM mal asignado y la probabilidad de la segunda puntuación, utilizando por tanto al mismo tiempo propiedades de distribuciones de péptidos individuales. Aunque la relación entre la calidad y la posición en el *ranking* es obvia puesto que la distribución promedio o *average* se construye ordenando los espectros según su mejor puntuación y por tanto la posición relativa en el *ranking* indica la fracción de espectros con una puntuación mejor, el análisis de las distribuciones individuales mediante el modelo teórico predice que al aumentar el número de

candidatos aumenta la precisión con que el ranking se estima a partir de la calidad. En otras palabras, que el método de la razón de probabilidades sería tanto más preciso cuanto mayor sea el número de candidatos utilizado, es decir, cuanto mayor sea el tamaño de las bases de datos utilizadas o más relajadas sean las condiciones con las que se lleva a cabo la búsqueda.

A pesar de que el concepto de la razón de probabilidades sea independiente de la calidad espectral, con anterioridad a esta tesis doctoral no se había evaluado el comportamiento de la razón de probabilidades en función del ancho de las distribuciones evaluadas. La evaluación conjunta de espectros que siguen distribuciones de la razón de probabilidades con una dispersión diferente – a modo de ejemplo, los espectros de carga 3 siguen una distribución de puntuaciones de la razón de probabilidades más ancha que los espectros de carga 2 – puede causar una subestimación de PSMs correctos. En este trabajo se propone una transformación teórica de las puntuaciones en función de la carga y la masa, y se estudió el comportamiento – especialmente en la zona de baja probabilidad – de las distribuciones promedio generadas con las puntuaciones transformadas por carga y masa frente a las distribuciones generadas con las puntuaciones iniciales, mostrando claramente una mayor uniformidad en el ancho de las distribuciones para las puntuaciones transformadas, lo que conlleva después en una mejora en el rendimiento de identificaciones positivas, sobre todo en la región de FDR baja (valores de FDR de 0 a 1%) y, por tanto, en la región de mejores identificaciones, lo que demuestra que esta "ecualización" de espectros según la carga y masa del ión parental ha sido adecuada, situando un número superior de identificaciones correctas en la región de mejores PSMs. Es importante notar que la transformación utilizada es una transformación teórica y, por tanto, no está sujeta a ningún tipo de ajuste experimental para cada conjunto de datos ni depende en ningún momento de ningún factor externo, función empírica o parámetro ajustable a decisión del investigador que utilice la transformación.

Ya que las distribuciones promedio de puntuaciones reflejan la distribución de calidad, éstas pueden ser diferentes cuando se analizan diferentes muestras. Los factores que pueden afectar a estas distribuciones son aquéllos que alteren la distribución de calidad de la muestra. Por ejemplo, incrementar la concentración de péptidos en la muestra incrementa por lo general la proporción de espectros con una buena fragmentación y por tanto la cantidad de espectros que tienden a producir una buena puntuación solamente por azar, por lo que la distribución de calidad se desplazará hacia la región de mejores valores. Si en cambio se modifica el tamaño de la base de datos – modificando el número de secuencias candidatas para cada espectro analizado – entonces la mayor parte de la distribución se desplazará a uno

u otro lado (dependiendo de si se aumenta o disminuye el número de candidatos), pero la forma de la distribución será la misma (D Lopez-Ferrer et al., 2004). Es una propiedad importante para la *pR*, ya que al ser ésta una razón de probabilidades, los desplazamientos de la distribución no afectan a la *pR*, con lo que se pueden utilizar distribuciones medias de puntuación generadas con conjuntos de datos muy grandes para poder estimar de una forma más precisa el factor *pR* en PSMs de experimentos más pequeños y que, generalmente, no alcanzan un número de espectros crítico tal que pueda generarse correctamente una distribución promedio de puntuaciones. Así, el factor *pR* puede ser utilizado como un parámetro muy preciso a nivel de espectro semejante a los estimados mediante el análisis de distribuciones de espectros individuales, con la clara ventaja en sencillez conceptual y tiempo de cómputo de que sólo necesita la información de la primera y segunda mejores puntuaciones.

Es generalmente reconocido, y los estudios sobre distribuciones que se realizan en (S Martinez-Bartolome et al., 2008) así lo corroboran, que las relaciones entre probabilidades promedio – razón de probabilidades – o puntuaciones umbral y tasas de error dependen en un grado muy alto del tamaño del experimento, del tamaño de la base de datos y de las condiciones de búsqueda utilizadas, por lo que su uso no puede ser extrapolado de datos obtenidos en otras muestras o buscados en condiciones diferentes. Por lo que es imposible establecer criterios fijos de identificación – como una puntuación umbral – de validez universal, aunque éstas sean prácticas habituales entre la comunidad científica(L Florens et al., 2002, AJ Link et al., 1999, J Peng et al., 2003, WJ Qian et al., 2005, MP Washburn et al., 2001). Cuando se analizan estadísticamente las puntuaciones de SEQUEST, los dos parámetros más utilizados son el Xcorr (o la mejor puntuación de correlación) y la diferencia relativa entre la mejor y la segunda mejor puntuación ΔC_n (JR Yates, 3rd et al., 1995, JR Yates, 3rd et al., 1995). Se asume por regla general que estos dos parámetros contienen la información más relevante sobre la identificación peptídica(A Keller et al., 2002, T Kislinger et al., 2003, D Lopez-Ferrer et al., 2004, RE Moore et al., 2002). Sin embargo, en su momento no estaba claro cuál era el método óptimo para analizar estadísticamente estos dos parámetros de forma conjunta, siendo el modelo matemático de Fernando Martín-Maroto el primero que predice mediante consideraciones puramente analíticas cuál es la forma óptima de considerar estos dos parámetros. El conjunto de este trabajo muestra en todo caso que el método *pR* obtiene un rendimiento superior que otros métodos previamente publicados(JE Elias,SP Gygi, 2007, A Keller et al., 2002, D Lopez-Ferrer et al., 2004) en el número de identificaciones positivas para

mismas tasas de error FDR. Este rendimiento superior es aún más evidente para la región de tasas de error más bajas (FDR < 1%).

La *pR* puede ser calculada de una manera extremadamente sencilla y aparamétrica, lo que impide la introducción de parámetros de ajuste que puedan introducir fuentes de error externas. Esto hace que sea un parámetro idóneo para su implementación en un proceso automático, sin necesidad de control humano. Es además un concepto completamente novedoso, derivado del análisis de consideraciones puramente analíticas, y mucho más simple de aplicar en la práctica que otras aproximaciones empíricas. Además, ofrece otras ventajas: al introducir una corrección interna para cada espectro, no es necesario clasificar los espectros de acuerdo a la carga y masa de su ión precursor(A Keller et al., 2002, D Lopez-Ferrer et al., 2004) ni se deben elegir funciones matemáticas predefinidas o parámetro ajustable alguno(J Colinge et al., 2003, A Keller et al., 2002, D Lopez-Ferrer et al., 2004, RG Sadygov et al., 2004). Además la corrección de calidad introducida hace al método muy robusto con respecto al uso de distribuciones de calidad particulares y con respecto al método utilizado para calcular valores de FDR utilizando bases de datos señuelo.

Aplicación del punto isoeléctrico en la inferencia estadística

El uso de técnicas de separación de péptidos permite mejorar la inferencia estadística de las identificaciones, utilizando las propiedades de la separación, que en general representan métodos ortogonales en relación al utilizado en inferencia estadística basado en las búsquedas en bases de datos de los espectros MS/MS. Son por tanto técnicas ideales para reducir el espacio de candidatos analizados, tanto en los resultados de búsquedas de la base de datos objetivo – *target* – como de la base de datos señuelo – *decoy*. La reducción del espacio de candidatos depende exclusivamente de la resolución del sistema de separación.

La tecnología Off-Gel usada en este trabajo, introducida de forma relativamente reciente(P Horth et al., 2006), tiene la ventaja de ser altamente reproducible, de rápida manipulación (no necesita la extracción de los péptidos del gel) y, sobre todo, de ser idónea para el análisis masivo, habiéndose demostrado que es la aproximación de fraccionamiento de péptidos que permite el mayor rendimiento en experimentos de identificación de péptidos a gran escala (E Bonzon-Kulichenko et al., en revisión). En esta aproximación se utilizan tiras de poliacrilamida

conteniendo anfolitos inmovilizados (tiras “IPG”) que permiten un control muy preciso del pH local en cada punto de la tira. Este control local del pH permite usar el punto isoeléctrico de los péptidos como un parámetro adicional para mejorar la inferencia estadística.

La integración del parámetro de punto isoeléctrico se realiza con una aproximación teórica sencilla: se evalúa la probabilidad global como el producto de la probabilidad de que el péptido esté bien identificado por espectrometría de masas (un PSM correcto) por la probabilidad de que el punto isoeléctrico del péptido identificado pertenezca a la fracción correspondiente. En nuestro caso hemos observado empíricamente que el método óptimo para modelar esta última es usando una función cuadrada de probabilidad, de ancho ajustable. El método de cálculo de la resolución utilizado es iterativo, y se inicia desde el caso menos resolutivo posible (un rango de pH 1-14), que va disminuyendo en cada iteración, hasta maximizar el número de identificaciones positivas (con un FDR menor de 1%). Este procedimiento es sencillo y robusto y permite que el método sea conservativo, ya que en ningún caso se obtendrán menos identificaciones que las que se obtendrían si no se utilizara este parámetro como indicador adicional. Esta cualidad hace a nuestro método particularmente idóneo para ser usado de forma automática y desatendida por el usuario, en experimentos de identificación a escala masiva.

En este trabajo el nuevo método se aplicó a dos experimentos de identificación masiva usando dos rangos de pH diferentes; en uno la separación de los péptidos por IEF se realizó en un rango de pH de 4 -7 y en el otro, de pH 3-10. La mejora en la tasa de identificación fue superior cuando se utilizó la tira de menor rango de pH (4-7); este resultado es esperable, porque ya que en estas condiciones la reducción del espacio de secuencias candidatas es superior. Este resultado sugiere que el método podría funcionar óptimamente usando tiras de pH de rangos todavía más estrechos, por ejemplo entre 4 y 5; sin embargo estas tiras producen una separación excesiva de los péptidos, la mayoría de los cuales se focaliza en los extremos de la tira, lo que resulta en una pérdida de rendimiento, por lo que no resultan útiles en la práctica.

Estimas de la tasa de error (FDR) usando bases de datos señuelo

En este trabajo se comparan dos estimaciones de FDR (la aproximación de bases de datos separadas, o método SD, y la aproximación de bases de datos concatenadas, o método CD) y se propone otro método de estimación que utiliza conceptos de los dos métodos anteriores, al que se nombra método *target-decoy*, o método TD.

Observando los métodos SD y CD no se puede concluir de forma trivial que el método CD sea mejor que el SD, ya que la mejora que introduce el método CD en la precisión de la estimación de PSM falsos positivos la realiza aumentando de forma artificial la población de referencia incluyendo en ella los PSM de la región **do** (*decoy only*), que corresponden a secuencias peptídicas que no existen en la naturaleza. Como las dos estimaciones de FDR se realizan sobre poblaciones de referencia diferentes no podemos afirmar de forma concluyente que un FDR calculado en la población de referencia es peor que un FDR más preciso estimado en una población aumentada de forma artificial. Este problema puede resolverse utilizando como base la distribución completa: la población de verdaderos positivos es común en las poblaciones de referencia de los dos métodos anteriores, por lo que una estimación de FDR en la población de referencia correcta puede calcularse estimando los PSM falsos positivos utilizando las regiones **do** y **db** y utilizando la simetría de los PSM aleatorios debida a la equiprobabilidad de resultados de puntuación en las bases de datos señuelo y objetivo, tal y como proponemos con el método *target-decoy*.

Comparando las ecuaciones Ecuación R. 7 y Ecuación R. 9 es evidente que la estrategia de competición (que el método TD utiliza) es mejor que la estrategia de bases de datos separadas SD, ya que el método TD no sobreestima el número de PSM falsos positivos en el numerador.

En cuanto a los métodos CD y TD, los dos están basados en la misma estrategia de competición (o simetría diagonal) y deben por tanto tener la misma precisión, por lo que el método TD se puede ver como un refinamiento del método CD, teniendo en consideración que el método TD no sobreestima el número de PSM falsos positivos, por lo que calcula el FDR sobre la población de referencia correcta.

En otra perspectiva, la aplicación del método TD exige realizar las búsquedas en bases de datos separadas, ya que es necesario determinar el tamaño de la población de referencia correcta (en concreto, debe determinarse el número de casos en la región **db**), por lo que el

método TD puede considerarse también como una mejora del método SD en la que se explota la simetría a lo largo de la diagonal de la distribución conjunta para calcular el número de PSM falsos positivos.

Como se ha comentado, los métodos CD y SD usan poblaciones de referencia diferentes, y sus estimaciones de FDR no deberían ser comparadas directamente. Sin embargo, estas comparaciones se pueden encontrar en la literatura(JE Elias,SP Gygi, 2007, L Kall et al., 2008, DL Tabb, 2008). Es interesante comparar en un marco de igualdad, utilizando para ello el mismo umbral de puntuación para los tres métodos. Repasando las tres ecuaciones correspondientes a los métodos expuestos, y ya que el número de casos en la región **tb** es por lo general mayor o igual que en **db** y que el número de casos en **do** es siempre mayor que cero, podemos concluir que en la mayoría de casos $FDR_{SD} \geq FDR_{TD}$ y $FDR_{CD} \geq FDR_{TD}$, que es lo que ya habíamos concluido en los dos párrafos anteriores: el método TD es mejor que los otros dos métodos. La comparación entre los métodos SD y CD es más compleja, por la ya explicada diferencia entre sus poblaciones de referencia. Para hacer comparaciones correctas se puede establecer una estimación paralela, FDR_{CD2}, del FDR del método concatenado con el mismo número de verdaderos positivos que la aproximación SD. En pocas palabras, se puede apreciar que en el caso de las puntuaciones Xcorr de SEQUEST, que tienen una gran correlación entre las puntuaciones en las bases de datos señuelo y objetivo, el método SD se ve penalizado ya que la región **tb** tiende a tener muchos PSM por esta correlación. En cuanto a las puntuaciones probabilísticas (pR y MASCOT), teniendo en cuenta FDR_{CD2}, el método CD es ligeramente mejor que el método SD utilizando MASCOT, pero en el caso de las puntuaciones pR ambos métodos tienen un comportamiento muy parecido, ya que la población en la región **tb** es mucho menor que con otras puntuaciones, y la penalización al método SD es por tanto menor. Con las puntuaciones de MASCOT, los tres métodos ofrecen un comportamiento muy similar para valores de FDR bajos.

En resumen, el método propuesto en este trabajo siempre ofrecerá un rendimiento superior (o igual) que los otros dos métodos comparados. Es un método que aprovecha las mejores propiedades de los dos métodos anteriores, ya que utiliza la población de referencia correcta (como el método SD) y utiliza el concepto de simetría debida a la equiprobabilidad de resultados (como el método CD).

Avances en Proteómica Cuantitativa: un modelo universal para la técnica de marcaje isotópico estable

Analizar un experimento a gran escala de proteómica cuantitativa es un desafío complejo en el que se necesita manejar herramientas estadísticas adecuadas. En nuestro laboratorio nos planteamos realizar este tipo de análisis con marcaje isotópico ^{18}O, y para ello realizamos pruebas con métodos estadísticos clásicos que no ofrecieron un marco adecuado para el análisis, produciendo falsos positivos. Este resultado, confirmado en múltiples ocasiones posteriores, motivó el desarrolló de un nuevo modelo estadístico apropiado para proteómica cuantitativa válido para cualquier tipo de marcaje isotópico y aparato de medida (espectrómetro de masas) utilizado.

El modelo estadístico

El modelo estadístico presentado en esta tesis se caracteriza por tres rasgos fundamentales. En primer lugar, el modelo considera que no todas las medidas de cuantificaciones relativas se realizan con la misma precisión, por lo que no se puede asumir que la varianza sea homogénea en la población completa de medidas, y por ello es necesario clasificar los datos en subconjuntos de medidas que tengan la misma varianza. Esta clasificación se realiza utilizando el peso estadístico, que es un parámetro que mide la dispersión de la medida realizada en función de una serie de parámetros exclusivos de la medición (intensidad espectral y ajuste de la envoltura isotópica experimental a una curva teórica, en los casos en que se utilice un ajuste de este tipo), y que se han definido de forma específica para todas y cada una de las técnicas de marcaje isobárico y espectrómetros de masas analizados en este trabajo. En este trabajo, no sólo hemos demostrado que los subconjuntos de medidas teniendo pesos similares se comportan en todos y cada uno de los casos como distribuciones normales, sino también que es posible estimar la dispersión de cada uno de esos subconjuntos de medidas directamente a partir del peso estadístico. Ello tiene la importante repercusión práctica de que es posible asignar un error – o varianza --, debido exclusivamente a la forma de medida, a todas y cada una de las medidas obtenidas, a partir de su peso estadístico. Hay otros trabajos en los que se ha esbozado, si bien de forma muy

preliminar, este mismo concepto; por ejemplo, en un trabajo reciente, utilizando marcaje metabólico $^{14}N/^{15}N$, se ha descrito que la varianza de la cuantificación de péptidos no es homogénea, y que ésta correlaciona con la razón señal/ruido (C Pan et al., 2006, C Pan et al., 2006); esto concuerda con los resultados presentados en este trabajo, puesto que en todas las aproximaciones expuestas en este estudio los pesos estadísticos son proporcionales al cuadrado de la intensidad de la señal.

En segundo lugar, el modelo tiene en cuenta tres fuentes de error correspondientes a cada uno de los niveles de cuantificación: el error cometido durante la medida en el espectrómetro de masas, el error con que diferentes péptidos cuantifican la proteína de la que se generan, y la dispersión con que cada una de las proteína se desvía del valor promedio esperado. Para ello se utiliza un modelo jerárquico de efectos aleatorios (que en algunos casos contiene también un efecto fijo) y calculando las medias de péptido y proteína teniendo en cuenta las contribuciones al error de cada una de las tres fuentes. Las medias de péptido y proteína son calculadas como medias ponderadas usando como pesos estadísticos la inversa de las varianzas de los diferentes valores con los que la media es calculada. Esta forma de proceder, además de ser estadísticamente correcta, integra todos los resultados de una manera coherente e intuitiva: las mediciones más precisas (con una varianza menor) son más determinantes que las de baja calidad (con una varianza mayor). Además, se tiene en cuenta de forma natural el hecho de que las cuantificaciones en cada uno de los niveles superiores (péptido o proteína) son más fiables cuando se estiman a partir de dos o más cuantificaciones de los niveles inferiores (espectros o péptidos, respectivamente). Además, se asume que cada uno de los niveles de medida (espectro, péptido y proteína) tiene un error intrínseco de medida. En el caso de espectros, este error depende fundamentalmente del aparato de medición – el espectrómetro de masas-, en el caso de los péptidos se asume un error debido fundamentalmente a los tratamientos posdigestión (excepto en el caso de SILAC, donde el marcaje es previo a la digestión y, por tanto, este error debería ser despreciable, que es precisamente el resultado que se obtiene al aplicar el modelo) y en las proteínas que se cuantifican en el experimento se asume también un error intrínseco debido a la variabilidad biológica del modelo analizado y a las técnicas de extracción de proteína que se utilizaron. Estos errores intrínsecos son permanentes, y son añadidos a los errores que se cometen en los respectivos niveles inferiores. Esto hace que no sea posible un caso extremo en el que, por ejemplo, un péptido medido por un número muy grande de espectros tenga un error nulo (ya que el error en la medición de un péptido debido a las mediciones de sus espectros disminuye en función del número de espectros que tomen parte en la medida, pero por muy pequeño

que sea nunca evitará el error intrínseco cometido en la propia generación del péptido) y, por tanto, la medida a nivel de proteína sea excesivamente sesgada hacia el valor de este péptido. De manera que este modelo tiene en cuenta tanto la calidad de las medidas como el número de réplicas en cada nivel. Este sistema de pesos estadísticos que definen la calidad de la medida (que estadísticamente hablando son en realidad la inversa de su varianza) se ha usado recientemente en el campo de micro matrices *(microarrays)*, asignando bajos pesos estadísticos a aquellas matrices menos reproducibles, lo que se ha demostrado que incrementa la capacidad de detección de cambios de expresión a nivel de mRNA (ME Ritchie et al., 2006). El modelo además permite contestar a dos preguntas fundamentales en proteómica cuantitativa: ¿cuál es el número mínimo de péptidos para considerar fiable a la cuantificación? Y, ¿cuál es el mínimo cambio de expresión que puede considerarse estadísticamente significativo? Según nuestro modelo, no hay un número mínimo de péptidos necesario, ni un cambio mínimo: todo queda en función de la calidad individual de la cuantificación, y esta calidad viene expresada en función de la varianza asociada a la medida. Una cuantificación realizada con un solo péptido puede ser fiable si el error asociado a la medida de ésta es pequeño. Y del mismo modo, un cambio de expresión pequeño puede ser estadísticamente significativo si el error asociado a su cuantificación es pequeño (lo que se consigue con muchas medidas de buena calidad), y un cambio de expresión grande puede no ser significativo, si la calidad de la cuantificación es muy baja. Esto puede verse en figuras como las presentadas en el capítulo de resultados en las que se representan los pesos de proteína (w_q) frente las medidas corregidas de las cuantificaciones de proteína (x_q-x). Se puede argumentar que el cambio de expresión de una proteína es menos sensible a artefactos si se utiliza más de un péptido para su cuantificación, sin embargo, en nuestras manos, en el gran número de experimentos realizados con este modelo observamos que las cuantificaciones atípicas de péptido fuera de los límites permitidos por su varianza asociada constituyen un porcentaje mínimo despreciable del conjunto de cuantificaciones de péptidos (en torno a menos del 0.1%). Esto indica que en la práctica todos los péptidos de una proteína, prácticamente sin excepción son fiables para cuantificar su expresión relativa, cada uno dentro de los límites impuestos por su propia varianza. Y en este contexto en el que prácticamente no hay medidas atípicas resulta razonable considerar como válidas las proteínas cuantificadas con un único péptido en los métodos propuestos. En el caso de cuantificaciones de péptidos marcados mediante iTRAQ, se ha descrito en la literatura (M Bantscheff et al., 2008, SY Ow et al., 2009) que los péptidos co-eluidos e isobáricamente próximos, junto con el fondo de contaminantes de naturaleza peptídica que se detecta durante el análisis de proteomas complejos, efectos que no se detectan en el espectro, pueden alterar las cuantificaciones, causando una

subestimación de los cambios de expresión(SY Ow et al., 2009), por lo que los cambios de expresión de proteínas cuantificadas con un solo péptido deben considerarse con cautela.

Hemos detectado dos fuentes de error en las cuantificaciones de péptidos difícilmente controlables: la oxidación de metioninas y la digestión incompleta realizada por el enzima utilizado – típicamente tripsina, fenómenos que pueden ocurrir en proporción diferente en las muestras comparadas. Recientes mejoras en los protocolos de digestión y marcaje(E Bonzon-Kulichenko et al., en revisión) nos han permitido minimizar estos efectos, homogeneizando el comportamiento de las muestras tratadas. Es muy importante hacer notar que el desarrollo de este modelo estadístico, unido al desarrollo de un software capaz de representar rápidamente todos los parámetros utilizados por el modelo, permite controlar en cada proteoma analizado posibles problemas que hayan podido surgir en cualquiera de las etapas clave del procesamiento de muestra.

La tercera propiedad importante del modelo es la suposición de que las varianzas asociadas a las fuentes de error de cada uno de los tres niveles – espectro, péptido y proteína – son constantes y comunes para todo el experimento. Este punto de vista global contrasta con la aproximación típica realizada en micro matrices en estudios de transcriptómica, o en geles de electroforesis bidimensional en estudios clásicos de proteómica cuantitativa, en los que se utiliza de forma habitual el test local de la t de Student. En estos test se asume que cada medida a nivel de proteína tiene una varianza diferente, aunque se lleve a cabo el mismo número de determinaciones. En el campo de las micro matrices se ha descrito recientemente (DB Allison et al., 2006, P Baldi,AD Long, 2001, AA Fodor et al., 2007) la dificultad y los enormes errores que puede conllevar realizar estimaciones locales de la varianza de todos y cada uno de los genes individuales, y cuando la varianza de cada gen es desconocida, entonces tiene sentido considerar que todos los genes de una matriz provienen de una única distribución normal (AA Fodor et al., 2007). En nuestro caso podemos afirmar que la varianza es homogénea en cada uno de los tres niveles basándonos en las pruebas de normalidad realizadas en cada uno de éstos. En el caso de los espectros, todos los experimentos pasaron una prueba de normalidad local (es decir, en datos con el mismo peso estadístico) basada en un test de D'Agostino (RB D'Agostino, 1971). Por otro lado, se ha definido también una variable normalizada $z_{s,}$ que representada frente a su valor teórico $z_{s,th}$ permite detectar desviaciones de la normalidad (“gráfica de normalidad”) (RB D'Agostino et al., 1990). En todos los casos, las representaciones confirman el ajuste del conjunto completo de datos a una distribución normal única. Este tipo de representación también se ha utilizado a nivel de péptido

(utilizando la variable normalizada z_p) y a nivel de proteína (utilizando z_q), confirmando en todos y cada uno de los casos la normalidad de los conjuntos analizados. El número de cuantificaciones atípicas en cada uno de los niveles es, además, prácticamente despreciable, y esto no sería posible si la varianza de alguno de los niveles no fuese homogénea. Como este número de cuantificaciones atípicas es despreciable, no supone problema la eliminación de estos datos del análisis, ya que difícilmente restarán información importante del conjunto de datos. Aquí conviene indicar que, en claro contraste, la eliminación de cuantificaciones atípicas de espectros y de péptidos es una práctica habitual en proteómica cuantitativa utilizando por ejemplo el criterio de Dixon(XJ Li et al., 2003, MJ MacCoss et al., 2003, SK Park et al., 2008), sin embargo no existen estudios sobre el número de cuantificaciones atípicas observado o las razones por las que se eliminan esos datos aparentemente erróneos.

Comparativa entre las diversas aproximaciones experimentales

Una de las mayores ventajas de nuestro modelo estadístico es que permite analizar las fuentes de error por separado y ello es particularmente útil a la hora de comparar las prestaciones relativas de cada uno de los métodos de marcaje isobárico. La comparación entre marcajes isotópicos en términos de error cometido debe realizarse en los niveles que pueden verse afectados por la manipulación de muestra, durante el marcaje y una vez ésta es marcada.

Los métodos de marcaje pre-digestión como SILAC no se ven afectados por la manipulación de la muestra una vez que ésta se digiere. Por tanto, el error sistemático – varianza – cometido a nivel de péptido debe ser nulo. Consistentemente, en todas las muestras marcadas con SILAC analizadas se observa una varianza de péptido σ^2_P muy baja, que puede considerarse despreciable. En cuanto a los marcajes isotópicos de ^{18}O y iTRAQ, la varianza de péptido σ^2_P es muy similar, pero ligeramente superior en los proteomas marcados mediante iTRAQ. Aunque este ligero efecto no lo hemos analizado en detalle, una posibilidad que no podemos descartar es que se deba a reacciones secundarias ocurridas durante el marcaje, por ejemplo alquilaciones sobre otros nucleófilos distintos a las aminas N-terminales o a los grupos amino de las Lys, tales como His ó Tyr, que tengan lugar en un grado de extensión ligeramente diferente en las muestras que se comparan.

Por otro lado, el tipo de marcaje utilizado afecta a la hora de llevar a cabo las cuantificaciones; dicho de otro modo, al error cometido a nivel de espectro o de medida. En este aspecto se observa una diferencia clara entre los métodos que utilizan espectros tipo MS para cuantificar (^{18}O, SILAC) con respecto a métodos cuya cuantificación se realiza en espectros tipo MS^2 (iTRAQ), siendo los primeros en general más precisos (la varianza a nivel de espectro σ^2_S llega a ser diez veces menor que en iTRAQ). Entre las aproximaciones ^{18}O y SILAC, la diferencia de varianza es pequeña en baja resolución, siendo la aproximación de SILAC ligeramente más precisa, probablemente debido a que en el caso de marcaje con ^{18}O los grupos de picos isotópicos correspondientes a las especies ligera y pesada tienen un mayor solapamiento, ya que la distancia de marcado es menor (4 Da en ^{18}O frente a 6 Da en SILAC). Esto hace que los ajustes por sumas de cuadrados sean algo menos precisos en el caso de ^{18}O.

Otro factor importante a tener en cuenta al comparar diferentes aproximaciones de marcaje isotópico es el rendimiento de identificaciones logradas que puedan asociarse a un espectro cuantificable, que en última instancia es lo que determina la profundidad de análisis de la muestra. En este aspecto, es evidente que iTRAQ tiene la ventaja de que la cuantificación se realiza sobre el mismo espectro MS^2 que se utiliza para la identificación, obteniéndose así un número mayor de espectros MS^2. Esta ventaja, no obstante, debe manejarse con cuidado al utilizar cromatografía líquida acoplada a espectrometría de masas, ya la rapidez de ciclo (barrido completo + n espectros MS^2) será menor si se desea mejorar la precisión en la cuantificación a nivel de espectro (que como hemos comentado anteriormente es inferior a la de los otros métodos de medida) aumentando el tiempo que el equipo utiliza para llevar a cabo cada medida. En cuanto a los métodos basados en cuantificación en espectros MS, SILAC y ^{18}O en equipos de baja resolución, la rapidez de ciclo es menor que en iTRAQ porque hace falta obtener un espectro para cuantificar y otro para identificar. Ello conlleva un menor rendimiento de identificaciones. Este efecto no se refleja en los experimentos realizados en este trabajo porque la cantidad de muestra utilizada para llevar a cabo el marcaje iTRAQ está limitada por la disponibilidad de reactivo (suministrado en kits para marcar un máximo de 100 µg de proteína para cada uno de los cuatro canales), limitación inexistente en el caso de ^{18}O y SILAC (en los que se usaron 500 µg de muestra marcada). Otro factor a tener en cuenta es la pérdida debida a la manipulación de muestra; en este caso el mejor rendimiento se obtiene en el caso de SILAC, donde no hace falta someter a los péptidos al proceso de marcaje y desalado posterior. Ignorando este factor, hay que comentar que SILAC ofrece por lo general un mayor rendimiento de identificaciones (y por tanto de cuantificaciones) que ^{18}O, ya que en el segundo caso la menor distancia de marcado, mencionado en el párrafo anterior, provoca que en los

espectros MS^2 en los que se identifica el péptido co-existan iones de la especie ligera y pesada, dificultando la identificación de los motores de búsqueda. Finalmente, el rendimiento óptimo desde el punto de vista del número de cuantificaciones se obtiene en el equipo de alta resolución (LTQ-Orbitrap), donde la cuantificación se realiza en modo MS, pero directamente en el espectro de barrido completo y, además tiene la peculiaridad de que este espectro de alta resolución (llevado a cabo en el módulo Orbitrap de este equipo híbrido), puede generarse al mismo tiempo que los péptidos más intensos están siendo fragmentados en el módulo de trampa lineal, de manera que la cuantificación y la identificación se realizan de forma simultánea.

Propiedades de la resolución utilizada

La resolución del equipo de medida utilizada afecta principalmente a dos parámetros: el error cometido en la medida – la varianza a nivel de espectro σ^2_S – y el rendimiento de identificaciones logradas. En aproximaciones de marcaje en los que se cuantifica sobre un espectro MS (SILAC, ^{18}O), la varianza a nivel de espectro σ^2_S es ligeramente inferior al utilizar espectrómetros de alta resolución, pero en general la diferencia en este sentido es despreciable. En métodos que cuantifican sobre un espectro MS^2 (iTRAQ), en principio no debe haber una gran diferencia, ya que la precisión con la que se miden las relaciones masa/carga de los iones reporteros no afecta a la precisión con la que se miden las intensidades de éstos. En nuestro caso hemos realizado un experimento con un LC-ESI acoplado a LTQ con un modo de activación PQD (baja resolución) y un MALDI-TOF-TOF (alta resolución); el MALDI-TOF-TOF es más preciso en la cuantificación de iTRAQ, pero esta mejora no se debe a la mayor resolución del equipo, sino a que las intensidades de los iones reporteros conseguidas son mucho mayores, ya que la fragmentación en la trampa iónica se realiza con menos energía que la fragmentación cuadrupolar (el caso de MALDI-TOF-TOF). Además la trampa requiere de un modo de activación especial (PQD) para evitar la regla del tercio, y poder así abordar la región de fragmentos de masa pequeñas (necesaria para ver los iones reporteros de iTRAQ); este modo especial de activación produce una gran pérdida de rendimiento.

Software de proteómica cuantitativa

El desarrollo de software de proteómica cuantitativa ha aumentado de forma considerable en los últimos cuatro años, lo que en principio debe permitir al investigador elegir entre un número aceptable de aplicaciones de software con las que poder analizar sus resultados. Sin embargo, dado que en proteómica cuantitativa se utilizan diversas aproximaciones (iTRAQ, ICAT, SILAC, ^{18}O, TMT, label-free) y diversos espectrómetros de masas (de baja, media y alta resolución con diferentes modos de fragmentación) que exportan sus datos en formatos también diversos, en general el número de programas que pueden utilizarse para una combinación concreta de todos estos factores anteriores es muy limitado. Esto obliga a la comunidad científica a conocer el funcionamiento de múltiples herramientas informáticas que en muchas ocasiones se utilizarán una única vez. Además, la comparación de resultados obtenidos por diferentes herramientas de software es complicada. En la Tabla D.1 se muestra una perspectiva del software que puede encontrarse actualmente en proteómica cuantitativa. Tras la división clara que puede hacerse entre aquellos programas dedicados a aproximaciones basadas en marcaje isotópico (JS Andersen et al., 2003, AM Boehm et al., 2007, J Cox,M Mann, 2008, DK Han et al., 2001, XJ Li et al., 2003, WT Lin et al., 2006, MJ MacCoss et al., 2003, CJ Mason et al., 2007, M Palmblad et al., 2008, C Pan et al., 2006, G Wang et al., 2006) y aquellos dedicados a aproximaciones label-free (JC Braisted et al., 2008, Z Khan et al., 2009) (ya que existen pocas herramientas actualmente que sean capaces de cuantificar tanto en label-free como en marcaje isotópico (D Bouyssie et al., 2007)) existen otras subdivisiones importantes: no todas las herramientas de cuantificación pueden utilizar datos de cualquier espectrómetro de masas (apenas aquellos que utilizan (JC Braisted et al., 2008, P Lu et al., 2007, M Palmblad et al., 2008) el Trans Proteomics Pipeline (TPP) o están integrados en él (DK Han et al., 2001, XJ Li et al., 2003), o aquellos que utilizan (Z Khan et al., 2009) estándares como mzXML) y cualquier tipo de resolución. Además pocos programas integran datos de cualquier tipo de motor de búsqueda: el universo de los motores de búsqueda en este tipo de programas queda prácticamente reducido al uso de MASCOT o SEQUEST, con contadas excepciones de programas como Quant (AM Boehm et al., 2007), APEX (JC Braisted et al., 2008) y maxQuant (M Palmblad et al., 2008) que interpretan datos de varios motores de búsqueda, o Proteome Discoverer (Thermo), que además unifica todos los resultados de varias búsquedas realizadas con diferentes motores.

Publicación	Software	Metodos	Resolucion o MS	Buscador	Estadística
(P Lu et al., 2007), (JC Braisted et al., 2008)	APEX	spectralCounting	cualquiera (TPP)	MASCOT, SEQUEST (TPP)	No comparable, ya que es label-free
(XJ Li et al., 2003)	ASAPratio (TPP)	ICAT, cICAT, SILAC	cualquiera (TPP)	TPP	utiliza el Interact de TPP para hacer el rolling-up, pero no tiene en cuenta los valores de cuantificación
no publicado	Libra (TPP)	iTRAQ			
no publicado	pepXML2Excel	Exportador de resultados			
(DK Han et al., 2001)	XPRESS (TPP)	ICAT	baja resolución (datos exportados a TPP)		
(J Cox,M Mann, 2008)	MaxQuant	SILAC	Orbitrap, FT (Alta resolución)	cualquiera	Rolling-up peptido-proteína por redundancia peptidica. Cuantificacion por criterio FDR Benjamini-Hochberg.
(D Bouyssie et al., 2007)	MFPaQ	Label-free,SILAC,ICAT	no especificado (importa datos .mgf)	Mascot	Criterios de exclusión basados en umbrales. Incluye un tratamiento de péptidos identificados en fracciones consecutivas de un gel.
(JS Andersen et al., 2003)	MSQuant	SILAC, 15N (con ayuda de otro SW)	Orbitrap, FT (Alta resolución)	Mascot	Se calcula un X^2, y el rolling-up se hace con medianas de los X^2 de los péptidos
no publicado	MS-Spectre	Label-free	alta resolución (datos importados: mzML, mzXML, MZData)	no integrado	No comparable, ya que es label-free
(WT Lin et al., 2006)	Multi-Q	iTRAQ	cualquiera	cualquiera (preferente: SEQUEST (pRatio))	Su rolling-up no tiene en cuenta el error sistemático cometido en cada nivel
(M Palmblad et al., 2008)	muxQuant	15N	FTICR	TPP	uso de medianas para rolling-up
(C Pan et al., 2006)	ProRata	15N, 13C, SILAC, ICAT, 18O	Thermo (LCQ-DECA,LTQ,LTQ-FT)	SEQUEST	Maximum likelihood. Basado en pesos estadísticos proporcional a SN ratio. No tiene en cuenta error sistemático cometido en cada nivel.No controla la eficiencia.
(Z Khan et al., 2009)	PVIEW	label-free (XIC)	alta resolución	XITandem (y posiblemente otros, mzXML)	No comparable, ya que es label-free
(AM Boehm et al., 2007)	Quant	iTRAQ	cualquiera	MASCOT, SEQUEST	No hace realmente rolling-up hasta proteína. Cuantifica péptidos.
(G Wang et al., 2006)	QUIL	cICAT,SILAC,18O	Thermo (LCQ-DECA,LTQ,LTQ-FT)	SEQUEST	Detección automática de anchos de pico, eliminación de background. Prevención de coeluidos. No estima la eficiencia en ^{18}O.
(CJ Mason et al., 2007)	RAAMS	18O	FT	SEQUEST (no especificado, pero parece ser)	Estima de la eficiencia similar a la propuesta por (A. Ramos et al, 2007). Sin rolling-up.
(MJ MacCoss et al., 2003)	RelEx	15N, ICAT	Thermo (LCQ-DECA,LTQ,LTQ-FT)	SEQUEST	Tests de Dixon para identificar cuantificaciones atípicas. Utiliza p-values, no válido para alto rendimiento.
no publicado	Mascot Distiller	cualquiera basado en marcaje isotópico (iTRAQ, 18O, SILAC, ICAT, TMT...)	cualquiera (no todos probados por los autores)	MASCOT	sin estadística visible
no publicado	Proteome Discoverer (Thermo)	iTRAQ, TMT, SILAC, ICAT	Thermo (LCQ-DECA,LTQ,LTQ-FT)	SEQUEST, MASCOT, y Z-Core. Combinación de todos los motores de búsqueda, aunque FDR mal calculada.	uso de medianas para rolling-up. Diversos errores en desestimación manual.
(I Jorge et al., 2009)	QuiXoT	cualquiera basado en marcaje isotópico (iTRAQ, 18O, SILAC, ICAT, TMT...)	cualquiera (probado en LTQ, Orbitrap y MALDI-TOF-TOF)	cualquiera (SEQUEST y MASCOT son transformados directamente, otros son mapeados por cualquier tabla de datos)	(1)

(1) Ver texto.

Tabla D.1. Comparativa de Software de proteómica cuantitativa.

Lo más problemático del asunto es que cada herramienta utiliza sus propios algoritmos de estadística para el análisis de las cuantificaciones: cada uno de ellos realiza el *rolling-up* (interpretación de qué proteínas están identificadas y en qué proporción relativa de cada muestra se encuentran, en función de los péptidos cuantificados que a su vez se encuentran en función de los espectros cuantificados) de un modo diferente: unos (Proteome Discoverer, MSQuant (JS Andersen et al., 2003)) utilizan medianas para calcular los valores de péptido y

proteína, otros realizan un control sobre cuantificaciones atípicas (como RelEx (MJ MacCoss et al., 2003)) para mejorar el *rolling-up*, y otros ni siquiera realizan *rolling-up* (véase por ejemplo Quant (AM Boehm et al., 2007)). En algunos casos las medias se realizan de forma ponderada usando pesos estadísticos elegidos arbitrariamente, generalmente en función de la intensidad de pico. Pero, que nosotros sepamos, en ninguno de los casos se utilizan criterios estadísticos en los que se haya validado la hipótesis nula en experimentos a gran escala; al contrario, la suposición de normalidad global (que como hemos visto en este trabajo nunca se cumple) se asume de forma implícita en todos ellos, y, aunque en algunos casos el software permite evaluar la dispersión media de la medida (en términos de %CV, o coeficiente de variación), en ninguno de ellos se analizan las fuentes de error por separado.

La plataforma de software desarrollada durante este trabajo, QuiXoT, permite aunar todas las aproximaciones basadas en marcaje isotópico, analizadas mediante cualquier tipo de espectrómetro de masas y mediante cualquier motor de búsqueda (mediante diversas herramientas que consiguen importar datos de espectrometría de masas en un formato propio e importar datos sobre identificaciones peptídicas prácticamente en cualquier formato), de forma que todas las combinaciones posibles de aproximación y análisis de datos son tratadas en un marco de referencia común – un mismo modelo estadístico --, que reúne todas las características necesarias de un modelo que se enfrente a un análisis de proteómica cuantitativa: un correcto *rolling-up*, basado en parámetros estadísticos bien definidos, en un modelo de hipótesis nula que se ha demostrado que se cumple en todos los casos, que tiene en cuenta el error sistemático cometido en cada nivel, y con un criterio que identifica cuantificaciones atípicas en cada uno de los tres niveles de medida. Además, su alta adaptabilidad permite que puedan utilizarse en el análisis de muestras marcadas con reactivos que no se utilizan de forma general (tales como marcaje metabólico con ^{15}N, o mediante reactivos de grupo específicos, por ejemplo agentes alquilantes de grupos Cys en experimentos de proteómica redox).

Perspectivas de futuro

El conjunto formado por el modelo estadístico de cuantificación y la plataforma de software QuiXoT presentado en este trabajo deja abierto un marco en el que pueden explorarse nuevas tecnologías de proteómica cuantitativa. Recientemente laboratorios de espectrometría de masas especializados aplican una combinación de tecnología label-free con marcaje isotópico (marcando una muestra base contra la que se comparan las muestras restantes), y se necesitan desarrollos en software que permitan interpretar este tipo de resultados de forma automática. De forma similar se pueden combinar técnicas de MRM con marcaje mediante ^{18}O. Otra gran meta que debe alcanzar la proteómica cuantitativa es la integración de la biología de sistemas que permita discernir cambios de expresión de grupos de proteínas que realicen una misma función, de forma que aumente nuestra comprensión sobre los cambios producidos a nivel celular. Ello sería posible introduciendo niveles superiores en el modelo estadístico jerárquico desarrollado en este trabajo, de manera que la integración de la información se haría en un mismo marco estadístico. Pero esta meta sólo es plausible si se asienta en unas bases sólidas y seguras, usando un modelo estadístico universal y validado en la práctica, como el descrito en este trabajo, que acople correctamente todos los niveles inferiores (proteína, péptido, espectro…).

Conclusiones

1. El rendimiento del método de la razón de probabilidades en experimentos de identificación de proteínas a gran escala a partir de espectros de fragmentación mejora cuando se introduce una corrección de dispersión local que depende de la carga y de la masa de los péptidos y cuando se tiene en cuenta el comportamiento de los péptidos durante el fraccionamiento por isoelectroenfoque.

2. La razón de probabilidades constituye un método que por su sencillez, robustez, ausencia de parámetros ajustables y de funciones empíricas y mayor sensibilidad que los métodos actuales, es particularmente idóneo para la automatización del proceso de identificación de péptidos en este tipo de experimentos a gran escala.

3. La tasa de error en experimentos de identificación de péptidos a gran escala puede estimarse usando bases de datos objetivo y señuelo de forma separada, aprovechando el comportamiento equiprobable de las asignaciones aleatorias y usando la estrategia de competición entre bases de datos, de forma más sensible que los métodos usados actualmente.

4. La dispersión de los datos de expresión diferencial obtenidos mediante marcaje isotópico y espectrometría de masas parece describirse correctamente y de forma general mediante un modelo jerárquico de hipótesis nula que tiene en cuenta la distribución aleatoria del error cometido en la interpretación del espectro, la detección de la señal por el equipo, la preparación de los péptidos a partir de sus proteínas respectivas y la preparación de la muestra de proteínas.

5. Los parámetros que describen el modelo de hipótesis nula pueden obtenerse de forma suficientemente precisa a partir del análisis estadístico de los datos obtenidos en un experimento de cuantificación diferencial de proteínas a gran escala. El modelo jerárquico permite la detección eficaz de artefactos introducidos durante el experimento y de cambios de expresión significativos.

6. El modelo jerárquico de hipótesis nula permite la comparación directa de las fuentes de error de los datos de expresión diferencial obtenidos mediante los diferentes métodos, y ofrece un marco estadístico común para el análisis de los resultados.

7. Todos los modelos y algoritmos desarrollados se han implementado dentro de una plataforma informática llamada QuiXoT, que permite el análisis rápido, riguroso y semiautomático de los resultados de cualquier experimento de expresión diferencial a gran escala mediante marcaje isotópico estable.

Bibliografía

[1] R Aebersold y S Patterson. Proteins: analysis and design, Academic Press, New York, 1998.
[2] DB Allison, X Cui, GP Page y M Sabripour. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55-65.
[3] SF Altschul, W Gish, W Miller, EW Myers y DJ Lipman. Basic local alignment search tool. J Mol Biol 1990;215:403-10.
[4] JS Andersen, CJ Wilkinson, T Mayor, P Mortensen, EA Nigg y M Mann. Proteomic characterization of the human centrosome by protein correlation profiling. Nature 2003;426:570-4.
[5] DC Anderson, W Li, DG Payan y WS Noble. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J Proteome Res 2003;2:137-46.
[6] AL Armesilla, E Lorenzo, P Gomez del Arco, S Martinez-Martinez, A Alfranca y JM Redondo. Vascular endothelial growth factor activates nuclear factor of activated T cells in human endothelial cells: a role for tissue factor gene expression. Mol Cell Biol 1999;19:2032-43.
[7] P Baldi y AD Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 2001;17:509-19.
[8] M Bantscheff, M Boesche, D Eberhard, T Matthieson, G Sweetman y B Kuster. Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. Mol Cell Proteomics 2008;7:1702-13.
[9] M Bantscheff, M Schirle, G Sweetman, J Rick y B Kuster. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 2007;389:1017-31.
[10] SC Bendall, C Hughes, MH Stewart, B Doble, M Bhatia y GA Lajoie. Prevention of amino acid conversion in SILAC experiments with embryonic stem cells. Mol Cell Proteomics 2008;7:1587-97.
[11] M Bern, D Goldberg, WH McDonald y JR Yates, 3rd. Automatic quality assessment of peptide tandem mass spectra. Bioinformatics 2004;20 Suppl 1:i49-54.
[12] AM Boehm, S Putz, D Altenhofer, A Sickmann y M Falk. Precise protein quantification based on peptide quantification using iTRAQ. BMC Bioinformatics 2007;8:214.
[13] E Bonzon-Kulichenko, D Pérez-Hernández, E Núñez, P Martinez-Acedo, P Navarro, M Trevisan*, et al.* A robust method for quantitative high-throughput analysis of proteomes by 18O labeling. Mol Cell Proteomics en revisión.
[14] D Bouyssie, A Gonzalez de Peredo, E Mouton, R Albigot, L Roussel, N Ortega*, et al.* Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. Mol Cell Proteomics 2007;6:1621-37.
[15] JC Braisted, S Kuntumalla, C Vogel, EM Marcotte, AR Rodrigues, R Wang*, et al.* The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results. BMC Bioinformatics 2008;9:529.
[16] H Choi y AI Nesvizhskii. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res 2008;7:47-50.
[17] J Colinge, A Masselot, M Giron, T Dessingy y J Magnin. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 2003;3:1454-63.
[18] J Cox y M Mann. Is proteomics the new genomics? Cell 2007;130:395-8.
[19] J Cox y M Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008;26:1367-72.
[20] RB D'Agostino. An Omnibus Test of Normality for Moderate and Large Size Samples. Biometrika 1971;58:341-48.

[21] RB D'Agostino, A Belanger y RBJ D'Agostino. A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician 1990;44:316-21.

[22] LM de Godoy, JV Olsen, J Cox, ML Nielsen, NC Hubner, F Frohlich*, et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 2008;455:1251-4.

[23] JE Elias y SP Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4:207-14.

[24] D Fenyo y RC Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 2003;75:768-74.

[25] J Fievet, C Dillmann, G Lagniel, M Davanture, L Negroni, J Labarre*, et al.* Assessing factors for reliable quantitative proteomics based on two-dimensional gel electrophoresis. Proteomics 2004;4:1939-49.

[26] M Fitzgibbon, Q Li y M McIntosh. Modes of inference for evaluating the confidence of peptide identifications. J Proteome Res 2008;7:35-9.

[27] L Florens, MP Washburn, JD Raine, RM Anthony, M Grainger, JD Haynes*, et al.* A proteomic view of the Plasmodium falciparum life cycle. Nature 2002;419:520-6.

[28] AA Fodor, TL Tickle y C Richardson. Towards the uniform distribution of null P values on Affymetrix microarrays. Genome Biol 2007;8:R69.

[29] E Gamma. Design patterns : elements of reusable object-oriented software, Addison-Wesley, Reading, Mass., 1995.

[30] CS Gan, PK Chong, TK Pham y PC Wright. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). J Proteome Res 2007;6:821-7.

[31] LY Geer, SP Markey, JA Kowalak, L Wagner, M Xu, DM Maynard*, et al.* Open mass spectrometry search algorithm. J Proteome Res 2004;3:958-64.

[32] SP Gygi, GL Corthals, Y Zhang, Y Rochon y R Aebersold. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci U S A 2000;97:9390-5.

[33] SP Gygi, B Rist, SA Gerber, F Turecek, MH Gelb y R Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 1999;17:994-9.

[34] DK Han, J Eng, H Zhou y R Aebersold. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol 2001;19:946-51.

[35] WJ Henzel, TM Billeci, JT Stults, SC Wong, C Grimley y C Watanabe. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc Natl Acad Sci U S A 1993;90:5011-5.

[36] ML Hernaez, C Gil, J Pla y C Nombela. Induced expression of the Candida albicans multidrug resistance gene CDR1 in response to fluconazole and other antifungals. Yeast 1998;14:517-26.

[37] P Horth, CA Miller, T Preckel y C Wenz. Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. Mol Cell Proteomics 2006;5:1968-74.

[38] Y Hu, G Wang, GY Chen, X Fu y SQ Yao. Proteome analysis of Saccharomyces cerevisiae under metal stress by two-dimensional differential gel electrophoresis. Electrophoresis 2003;24:1458-70.

[39] P James, M Quadroni, E Carafoli y G Gonnet. Protein identification by mass profile fingerprinting. Biochem Biophys Res Commun 1993;195:58-64.

[40] I Jorge, EM Casas, M Villar, I Ortega-Perez, D Lopez-Ferrer, A Martinez-Ruiz*, et al.* High-sensitivity analysis of specific peptides in complex samples by selected MS/MS ion monitoring and linear ion trap mass spectrometry: application to biological studies. J Mass Spectrom 2007;42:1391-403.

[41] I Jorge, P Navarro, P Martinez-Acedo, E Nunez, H Serrano, A Alfranca*, et al.* Statistical model to analyze quantitative proteomics data obtained by 18O/16O labeling and linear ion trap mass spectrometry: application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. Mol Cell Proteomics 2009;8:1130-49.

[42] L Kall, JD Storey, MJ MacCoss y WS Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 2008;7:29-34.

[43] NA Karp y KS Lilley. Design and analysis issues in quantitative proteomics studies. Proteomics 2007;7 Suppl 1:42-50.

[44] NA Karp, PS McCormick, MR Russell y KS Lilley. Experimental and statistical considerations to avoid false conclusions in proteomics studies using differential in-gel electrophoresis. Mol Cell Proteomics 2007;6:1354-64.

[45] A Keller, AI Nesvizhskii, E Kolker y R Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383-92.

[46] Z Khan, JS Bloom, BA Garcia, M Singh y L Kruglyak. Protein quantification across hundreds of experimental conditions. Proc Natl Acad Sci U S A 2009;106:15544-8.

[47] DS Kirkpatrick, SA Gerber y SP Gygi. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. Methods 2005;35:265-73.

[48] T Kislinger, K Rahman, D Radulovic, B Cox, J Rossant y A Emili. PRISM, a generic large scale proteomic investigation strategy for mammals. Mol Cell Proteomics 2003;2:96-106.

[49] V Lanquar, L Kuhn, F Lelievre, M Khafif, C Espagne, C Bruley*, et al.* 15N-metabolic labeling for comparative plasma membrane proteomics in Arabidopsis cells. Proteomics 2007;7:750-4.

[50] KW Lau, AR Jones, N Swainston, JA Siepen y SJ Hubbard. Capture and analysis of quantitative proteomic data. Proteomics 2007;7:2787-99.

[51] SS Li, J Bigler, JW Lampe, JD Potter y Z Feng. FDR-controlling testing procedures and sample size determination for microarrays. Stat Med 2005;24:2267-80.

[52] XJ Li, H Zhang, JA Ranish y R Aebersold. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal Chem 2003;75:6648-57.

[53] WT Lin, WN Hung, YH Yian, KP Wu, CL Han, YR Chen*, et al.* Multi-Q: a fully automated tool for multiplexed protein quantitation. J Proteome Res 2006;5:2328-38.

[54] AJ Link, J Eng, DM Schieltz, E Carmack, GJ Mize, DR Morris*, et al.* Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol 1999;17:676-82.

[55] D Lopez-Ferrer, S Martinez-Bartolome, M Villar, M Campillos, F Martin-Maroto y J Vazquez. Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. Anal Chem 2004;76:6853-60.

[56] D Lopez-Ferrer, A Ramos-Fernandez, S Martinez-Bartolome, P Garcia-Ruiz y J Vazquez. Quantitative proteomics using 16O/18O labeling and linear ion trap mass spectrometry. Proteomics 2006;6 Suppl 1:S4-11.

[57] P Lu, C Vogel, R Wang, X Yao y EM Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 2007;25:117-24.

[58] MJ MacCoss, CC Wu, H Liu, R Sadygov y JR Yates, 3rd. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. Anal Chem 2003;75:6912-21.

[59] MJ MacCoss, CC Wu y JR Yates, 3rd. Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal Chem 2002;74:5593-9.

[60] M Mann, P Hojrup y P Roepstorff. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom 1993;22:338-45.

[61] S Martinez-Bartolome, P Navarro, F Martin-Maroto, D Lopez-Ferrer, A Ramos-Fernandez, M Villar*, et al.* Properties of average score distributions of SEQUEST: the probability ratio method. Mol Cell Proteomics 2008;7:1135-45.

[62] CJ Mason, TM Therneau, JE Eckel-Passow, KL Johnson, AL Oberg, JE Olson*, et al.* A method for automatically interpreting mass spectra of 18O-labeled isotopic clusters. Mol Cell Proteomics 2007;6:305-18.

[63] V Mayya, K Rezaul, YS Cong y D Han. Systematic comparison of a two-dimensional ion trap and a three-dimensional ion trap mass spectrometer in proteomics. Mol Cell Proteomics 2005;4:214-23.

[64] OA Mirgorodskaya, YP Kozmin, MI Titov, R Korner, CP Sonksen y P Roepstorff. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. Rapid Commun Mass Spectrom 2000;14:1226-32.

[65] RE Moore, MK Young y TD Lee. Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom 2002;13:378-86.

[66] E Mortz, PB O'Connor, P Roepstorff, NL Kelleher, TD Wood, FW McLafferty*, et al.* Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases. Proc Natl Acad Sci U S A 1996;93:8264-7.

[67] LN Mueller, MY Brusniak, DR Mani y R Aebersold. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 2008;7:51-61.

[68] CJ Nelson, EL Huttlin, AD Hegeman, AC Harms y MR Sussman. Implications of 15N-metabolic labeling for automated peptide identification in Arabidopsis thaliana. Proteomics 2007;7:1279-92.

[69] AI Nesvizhskii, FF Roos, J Grossmann, M Vogelzang, JS Eddes, W Gruissem*, et al.* Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics 2006;5:652-70.

[70] JV Olsen, B Macek, O Lange, A Makarov, S Horning y M Mann. Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods 2007;4:709-12.

[71] SE Ong, B Blagoev, I Kratchmarova, DB Kristensen, H Steen, A Pandey*, et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 2002;1:376-86.

[72] SY Ow, M Salim, J Noirel, C Evans, I Rehman y PC Wright. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". J Proteome Res 2009;8:5347-55.

[73] M Palmblad, DJ Mills y LV Bindschedler. Heat-shock response in Arabidopsis thaliana explored by multiplexed quantitative proteomics using differential metabolic labeling. J Proteome Res 2008;7:780-5.

[74] C Pan, G Kora, WH McDonald, DL Tabb, NC VerBerkmoes, GB Hurst*, et al.* ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. Anal Chem 2006;78:7121-31.

[75] C Pan, G Kora, DL Tabb, DA Pelletier, WH McDonald, GB Hurst*, et al.* Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. Anal Chem 2006;78:7110-20.

[76] DJ Pappin, P Hojrup y AJ Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. Curr Biol 1993;3:327-32.

[77] SK Park, JD Venable, T Xu y JR Yates, 3rd. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods 2008;5:319-22.

[78] J Peng, JE Elias, CC Thoreen, LJ Licklider y SP Gygi. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2003;2:43-50.

[79] WJ Qian, T Liu, ME Monroe, EF Strittmatter, JM Jacobs, LJ Kangas*, et al.* Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. J Proteome Res 2005;4:53-62.

[80] WJ Qian, ME Monroe, T Liu, JM Jacobs, GA Anderson, Y Shen*, et al.* Quantitative proteome analysis of human plasma following in vivo lipopolysaccharide administration using 16O/18O labeling and the accurate mass and time tag approach. Mol Cell Proteomics 2005;4:700-9.

[81] A Ramos-Fernandez, D Lopez-Ferrer y J Vazquez. Improved method for differential expression proteomics using trypsin-catalyzed 18O labeling with a correction for labeling efficiency. Mol Cell Proteomics 2007;6:1274-86.

[82] J Razumovskaya, V Olman, D Xu, EC Uberbacher, NC VerBerkmoes, RL Hettich*, et al.* A computational method for assessing peptide- identification reliability in tandem mass spectrometry analysis with SEQUEST. Proteomics 2004;4:961-9.

[83] ME Ritchie, D Diyagama, J Neilson, R van Laar, A Dobrovic, A Holloway*, et al.* Empirical array quality weights in the analysis of microarray data. BMC Bioinformatics 2006;7:261.

[84] P Roepstorff y J Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 1984;11:601.

[85] PL Ross, YN Huang, JN Marchese, B Williamson, K Parker, S Hattan*, et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3:1154-69.

[86] RG Sadygov, H Liu y JR Yates. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal Chem 2004;76:1664-71.

[87] RG Sadygov y JR Yates, 3rd. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem 2003;75:3792-8.

[88] M Schnolzer, P Jedrzejewski y WD Lehmann. Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. Electrophoresis 1996;17:945-53.

[89] IP Shadforth, TP Dunkley, KS Lilley y C Bessant. i-Tracker: for quantitative proteomics using iTRAQ. BMC Genomics 2005;6:145.

[90] A Shevchenko, ON Jensen, AV Podtelejnikov, F Sagliocco, M Wilm, O Vorm*, et al.* Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc Natl Acad Sci U S A 1996;93:14440-5.

[91] A Shevchenko, H Tomas, J Havlis, JV Olsen y M Mann. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 2006;1:2856-60.

[92] GB Smejkal, MH Robinson y A Lazarev. Comparison of fluorescent stains: relative photostability and differential staining of proteins in two-dimensional gels. Electrophoresis 2004;25:2511-9.

[93] JD Storey y R Tibshirani. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440-5.

[94] EF Strittmatter, LJ Kangas, K Petritis, HM Mottaz, GA Anderson, Y Shen*, et al.* Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. J Proteome Res 2004;3:760-9.

[95] DL Tabb. What's driving false discovery rates? J Proteome Res 2008;7:45-6.
[96] DL Tabb, MJ MacCoss, CC Wu, SD Anderson y JR Yates, 3rd. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. Anal Chem 2003;75:2470-7.
[97] A Thompson, J Schafer, K Kuhn, S Kienle, J Schwarz, G Schmidt*, et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 2003;75:1895-904.
[98] G Wang, WW Wu, T Pisitkun, JD Hoffert, MA Knepper y RF Shen. Automated quantification tool for high-throughput proteomics using stable isotope labeling and LC-MSn. Anal Chem 2006;78:5752-61.
[99] MP Washburn, D Wolters y JR Yates, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 2001;19:242-7.
[100] JR Wisniewski, A Zougman, N Nagaraj y M Mann. Universal sample preparation method for proteome analysis. Nat Methods 2009;6:359-62.
[101] JX Yan, AT Devenish, R Wait, T Stone, S Lewis y S Fowler. Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of Escherichia coli. Proteomics 2002;2:1682-98.
[102] X Yao, C Afonso y C Fenselau. Dissection of proteolytic 18O labeling: endoprotease-catalyzed 16O-to-18O exchange of truncated peptide substrates. J Proteome Res 2003;2:147-52.
[103] X Yao, A Freas, J Ramirez, PA Demirev y C Fenselau. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. Anal Chem 2001;73:2836-42.
[104] JR Yates, 3rd, JK Eng y AL McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem 1995;67:3202-10.
[105] JR Yates, 3rd, JK Eng, AL McCormack y D Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 1995;67:1426-36.
[106] JR Yates, 3rd, S Speicher, PR Griffin y T Hunkapiller. Peptide mass maps: a highly informative approach to protein identification. Anal Biochem 1993;214:397-408.

Anexos

Anexo A

1. Cálculo de dispersiones – varianzas – de espectros, péptidos y proteínas.

La estimación de las varianzas utilizadas en cada nivel (proteína, péptido y espectro) puede hacerse mediante una aproximación de máxima verosimilitud, sin embargo ésta es una aproximación sesgada, debido a que en un experimento típico de proteómica cuantitativa el número de péptidos por proteína y el número de espectros por péptido es muy variable, pero tiende a ser un número bajo, por lo que para estimar las dispersiones cometidas en las medidas se deben tener en cuenta correctamente los grados de libertad de cada dispersión medida. Se pueden calcular por dos métodos:

1.1 Método iterativo.

Asumamos que se analiza en un experimento un total de NQ proteínas, cada proteína es cuantificada por n_q péptidos diferentes, y cada péptido es medido por n_{qp} espectros. El número total de espectros y péptidos será, respectivamente,

$$NS = \sum_q \sum_p n_{qp}$$

ecuación A. 1

y

$$NP = \sum_q n_q$$

ecuación A. 2

Los componentes de varianza de espectro, péptido y proteína son calculados usando una población que contiene espectros de alta calidad donde los errores debidos al ajuste de la curva teórica pueden asumirse prácticamente nulos. Esta población se construye

seleccionando aquellos espectros con un peso estadístico v_{qps} superior a un valor por el que se observa que la población elegida tiene una varianza homogénea. Típicamente para ^{18}O en un equipo tipo trampa iónica (LTQ, Thermo) este valor es de $v_{qps} = 30$. Como se asume que en esta población la varianza de espectro es constante, todos los espectros contribuyen de la misma manera a las medias de los péptidos, y por tanto se pueden considerar medias no ponderadas:

$$x_{qp} = \frac{\sum_{s} x_{qps}}{n_{qp}}$$

ecuación A. 3

Sea:

$$SSS = \sum_{q}\sum_{p}\sum_{s}(x_{qps} - x_{qp})^2$$

ecuación A. 4

la suma de diferencias cuadráticas de todos los espectros con respecto a la media del péptido. Esta suma puede ser utilizada para estimar el componente de varianza a nivel de espectro mediante el método de ANOVA convencional:

$$\frac{E(SSS)}{NS - NP} = \sigma_{\mathrm{S}}^2$$

ecuación A. 5

Para calcular la varianza a nivel de péptido se utiliza un método iterativo. Se establece un valor inicial de varianza a nivel de péptido σ^2_{P}, por ejemplo cero, y se utiliza para un cálculo inicial de pesos estadísticos de péptido y las medias de proteína se calculan como medias ponderadas de sus péptidos:

$$w_{qp} = \frac{1}{\frac{\sigma_{\mathbf{S}}^2}{n_{qp}} + \sigma_{\mathbf{P}}^2}$$

ecuación A. 6

$$x_q = \frac{\sum_p w_{qp} x_{pq}}{\sum_p w_{qp}}$$

ecuación A. 7

La suma ponderada de diferencias cuadráticas entre medias de péptidos y sus correspondientes medias de proteína:

$$SSP = \sum_q \sum_p w_{qp} (x_{qp} - x_q)^2$$

ecuación A. 8

se utiliza para calcular la varianza a nivel de péptido usando la siguiente expresión:

$$\sigma_{\mathbf{P}}^2 = \frac{E(SSP)}{\sum_q W_q} - \frac{\sigma_{\mathbf{S}}^2 \sum_q \sum_p \frac{w_{qp}}{n_{qp}}}{\sum_q \sum_p w_{qp}}$$

ecuación A. 9

donde

$$W_q = \sum_p w_{qp} - \frac{\sum_p w_{qp}^2}{\sum_p w_{qp}}$$

ecuación A. 10

son los grados de libertad del sistema. Este factor asegura que la estimación local de la varianza de péptido calculada no esté sesgada cuando el número de péptidos utilizados para estimar la media de proteína es pequeño, como es habitual en este tipo de experimentos (figura R.16 en el texto principal). El valor obtenido de σ^2_P es utilizado entonces como estimación inicial para una repetición del proceso, y el proceso es iterado hasta que el valor de σ^2_P converge.

La varianza a nivel de proteína se calcula de un modo similar, escogiendo un valor inicial de σ^2_Q para calcular los pesos estadísticos de proteína y entonces estimar la supermedia como una media ponderada de proteínas:

$$w_q = \frac{1}{\frac{1}{\sum_p w_{qp}} + \sigma^2_{\mathbf{Q}}}$$

ecuación A. 11

$$x = \frac{\sum_q w_q x_q}{NQ}$$

ecuación A. 12

La suma ponderada de diferencias cuadráticas entre las medias de proteína y la supermedia

$$SSQ = \sum_q w_q (x_q - x)^2$$

ecuación A. 13

se utiliza para calcular la varianza a nivel de proteína mediante la expresión

$$\sigma_{\mathbf{Q}}^2 = \frac{E(SSQ)}{\sum_q w_q - \frac{\sum_q w_q^2}{\sum_q w_q}} - \frac{\sum_q w_q^2}{\sum_q w_q}$$

ecuación A. 14

Y el proceso se itera hasta alcanzar un valor estable de σ^2_Q.

1.2 Método robusto basado en medianas.

Las varianzas de proteína, péptido o espectro pueden calcularse también por una estimación robusta basada en la mediana de las desviaciones cuadráticas de las medidas, proporcional a la varianza:

$$\sigma^2 = (1.4826)^2 \cdot MED\left\{(x_i - x)^2 \cdot gdl\right\}$$

ecuación A. 15

donde gdl es una corrección local de los grados de libertad de cada subconjunto de espectros asociado a un péptido o de cada subconjunto de péptidos asociado a una proteína. Esta corrección se introduce porque al tratarse de medias ponderadas obtenidas a partir de un conjunto pequeño de medidas x_i , las estimas de varianza están fuertemente sesgadas, produciéndose una subestimación de sus valores. La corrección local tiene en cuenta la proporcionalidad entre los grados de libertad de la estima sesgada y de la no sesgada, de manera que gdl = n / (n-1), y se calcula localmente a nivel de cada medida, de manera que todas las medidas a nivel de espectro o a nivel de péptido pueden considerarse conjuntamente para hacer una única estimación de la varianza. El término 1.4826 es un factor de escala correspondiente a la inversa del 75ª percentil de una distribución normal con σ=1.

Y se puede desarrollar un método iterativo de estimación de la varianza σ, si normalizamos la ecuación anterior:

$$\frac{(1.4826)^2}{\sigma^2}\cdot MED\left\{(x_i-x)^2\cdot gdl\right\}=(1.4826)^2\cdot MED\left\{\frac{(x_i-x)^2\cdot gdl}{\sigma^2}\right\}=1$$

ecuación A. 16

Las medianas dependen estrictamente del orden, por lo que puede calcularse igualmente con los valores de la raíz cuadrada, que corresponden a los valores de la variable normalizada z_i:

$$(1.4826)^2\cdot MED\left\{\frac{(x_i-x)\cdot\sqrt{gdl}}{\sigma}\right\}=(1.4826)^2\cdot MED\{z_i\}=1$$

ecuación A. 17

De forma explícita para las varianzas de espectro, péptido y proteína:

$$Espectro:(1.4826)^2\cdot MED\left\{\frac{(x_{qps}-x_{qp})\cdot\sqrt{n_{pq}/(n_{pq}-1)}}{\sigma_S}\right\}=(1.4826)^2\cdot MED\{z_{qps}\}=1$$

$$Péptido:(1.4826)^2\cdot MED\left\{\frac{(x_{qp}-x_q)\cdot\sqrt{n_q/(n_q-1)}}{\sigma_P}\right\}=(1.4826)^2\cdot MED\{z_{qp}\}=1$$

$$\text{Pr}\,oteína:(1.4826)^2\cdot MED\left\{\frac{(x_q-x)}{\sigma_Q}\right\}=(1.4826)^2\cdot MED\{z_q\}=1$$

ecuación A. 18

Con lo que se puede realizar una estimación de las varianzas mediante un método iterativo variando la σ_i de interés, buscando el valor de la función normalizada más próximo a uno. Este método es resistente a la presencia de medidas desviadas atípicamente, ya que se calcula con valores de medianas, y es una estima no sesgada, ya que está corregida localmente por los grados de libertad de cada subconjunto de medidas.

*2. Estimación del factor Φ por la aproximación de máxima verosimilitud (*maximum likelihood*)*

La ecuación general para el modelo estadístico de un experimento de marcaje isotópico usando SILAC viene dada por

$$x_{qps} = \mu + \rho_q + \beta_{qp} + \xi_{qps}$$

ecuación A. 19

Donde la única diferencia con el modelo general es que el error a nivel de péptido se distribuye normalmente de acuerdo a $\beta_{qp} \sim N(r_{qp} \times \phi, \sigma^2_P)$, siendo r_{qp} el número de prolinas que contiene el péptido p perteneciente a la proteína q y g=2^(- ϕ) es la proporción de prolinas que no han sufrido marcaje metabólico a partir de arginina. Como se ha descrito en el texto principal, los errores ρ_q y ε_{qps} también se distribuyen normalmente con varianza σ^2_Q y $\sigma^2_S + k/v_{qps}$, respectivamente.

Puesto que el modelo supone que todos los errores se distribuyen normalmente, la función de verosimilitud L correspondiente a este modelo es el productorio de las distribuciones gaussianas alrededor de todas las medidas a los tres niveles:

$$L\left(\vec{\theta}\middle|\vec{X}\right) = L\left(k, \sigma_S^2, \sigma_P^2, \sigma_Q^2, \phi \middle| x_{qps}, x_{qp}, w_{qp}, r_{qp}\right) = \prod_q \prod_p \prod_s \frac{\sqrt{w_{qps}}}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} w_{qps} \left(x_{qps} - x_{qp}\right)^2 \right) \cdot$$
$$\cdot \prod_q \prod_p \frac{\sqrt{w_{qp}}}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} w_{qp} \left(x_{qp} - r_{qp}\phi - x_q\right)^2 \right) \cdot$$
$$\prod_q \frac{\sqrt{w_q}}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} w_q \left(x_q - x\right)^2 \right)$$

ecuación A. 20

El valor de ϕ que mejor describe los resultados experimentales de acuerdo al modelo estadístico se puede calcular por el método de la máxima verosimilitud, según el cual se estima el valor de ϕ como aquél que maximiza la función L. Como el máximo de L corresponde también al máximo de su logaritmo, el valor de ϕ tiene que ser tal que se verifique

$$\frac{\partial \ln L}{\partial \phi} = 0$$

ecuación A. 21

Desarrollando la derivada del logaritmo de L e igualando a cero se obtiene

$$\sum_q \sum_p w_{qp} \left(x_{qp} - r_{qp}\phi - x_q \right) \cdot r_{qp} = 0$$

ecuación A. 22

Y despejando ϕ

$$\phi = \frac{\sum_q \sum_p w_{qp} \left(x_q - x_{qp} \right) \cdot r_{qp}}{\sum_q \sum_p w_{qp} \cdot r_{qp}^2}$$

ecuación A. 23

que es la estima de máxima verosimilitud del parámetro ϕ. Nótese que ésta es una estima no sesgada y, por tanto, centrada correctamente en el valor esperado. Sin embargo la aplicación de este método de máxima verosimilitud al cálculo de las varianzas daría estimas fuertemente sesgadas debido a que el número de medidas por péptido y el número de péptidos por proteínas son valores muy próximos a la unidad. Por esta razón las varianzas del modelo se calculan de otra manera, detallada más adelante en este anexo.

Cambios de expresión observados en los experimentos de generalización del modelo estadístico.

Anexo A. Cambios de expresión estadísticamente significativos encontrados en los experimentos de generalización del modelo estadístico.

Proteínas cuantificadas con un valor de FDR_q inferior al 5%.

^{18}O LTQ — **A vs B**

#acceso	FASTAProteinDescription	Xq-X	SD_Xq	Wq	Zq	FDRq	pep_per_protein
P38143	sp|P38143|GPX2_YEAST Glutathione peroxidase 2 OS=Saccharo	-1.48601327	0.15569994	41.2499224	-9.544084	0	2
Q04120	sp|Q04120|TSA2_YEAST Peroxiredoxin TSA2 OS=Saccharomyces	-2.53343817	0.18890802	28.0220039	-13.41096	0	1
P40989	sp|P40989|FKS2_YEAST 1,3-beta-glucan synthase component G	1.5526305	0.18858258	28.1188032	8.233159	1.843E-13	1
P38286	sp|P38286|MKAR_YEAST 3-ketoacyl-CoA reductase OS=Sacchar	1.19398665	0.21171453	22.3099537	5.639606	3.5366E-06	1
P22803	sp|P22803|TRX2_YEAST Thioredoxin-2 OS=Saccharomyces cere	-0.75347955	0.13910019	51.6826252	-5.416812	1.2589E-05	3
P00431	sp|P00431|CCPR_YEAST Cytochrome c peroxidase, mitochondri	-0.7567865	0.15599573	41.0936363	-4.851328	0.00025447	2
Q12068	sp|Q12068|GRE2_YEAST NADPH-dependent methylglyoxal redu	-0.59672197	0.16293911	37.665986	-3.662239	0.02964529	2
P29509	sp|P29509|TRXB1_YEAST Thioredoxin reductase 1 OS=Saccharo	-0.58328521	0.15936028	39.3767491	-3.660167	0.02988606	2

^{18}O Orbitrap — **A vs B**

#acceso	FASTAProteinDescription	Xq-X	SD_Xq	Wq	Zq	FDRq	pep_per_protein
Q12177	sp|Q12177|YL056_YEAST Uncharacterized protein YLL056C OS=	-2.84107726	0.31318653	10.1951553	-9.071518	0	1
P41816	sp|P41816|OYE3_YEAST NADPH dehydrogenase 3 OS=Saccharo	-3.15083651	0.23397573	18.2666278	-13.46651	0	3
Q04120	sp|Q04120|TSA2_YEAST Peroxiredoxin TSA2 OS=Saccharomyces	-2.89004345	0.21410172	21.8152253	-13.49846	0	3
P38143	sp|P38143|GPX2_YEAST Glutathione peroxidase 2 OS=Saccharo	-1.68229988	0.21176921	22.2984348	-7.944025	9.93E-13	3
P00431	sp|P00431|CCPR_YEAST Cytochrome c peroxidase, mitochondri	-1.24704766	0.202204	24.4579766	-6.167275	3.45E-07	5
P53172	sp|P53172|SDS23_YEAST Protein SDS23 OS=Saccharomyces cer	1.56955631	0.26975914	13.7419281	5.818362	1.48E-06	1
P28007	sp|P28007|GAR1_YEAST H/ACA ribonucleoprotein complex sub	-1.48646258	0.26850284	13.8708235	-5.536115	7.68E-06	1
Q01574	sp|Q01574|ACS1_YEAST Acetyl-coenzyme A synthetase 1 OS=Sa	1.44475886	0.2726278	13.454257	5.299382	2.89E-05	1
Q12068	sp|Q12068|GRE2_YEAST NADPH-dependent methylglyoxal redu	-0.96300448	0.19627961	25.9567094	-4.906289	0.00015366	6
P42846	sp|P42846|KRI1_YEAST Protein KRI1 OS=Saccharomyces cerevis	-1.3646842	0.28456193	12.3494151	-4.795737	0.00026833	1
P29295	sp|P29295|HRR25_YEAST Casein kinase I homolog HRR25 OS=Sa	1.11573727	0.24078713	17.2477895	4.633708	0.00044597	2
P53741	sp|P53741|BRE5_YEAST UBP3-associated protein BRE5 OS=Sacc	-1.38208588	0.29645829	11.37818	-4.661991	0.00051846	1
Q12443	sp|Q12443|RTN2_YEAST Reticulon-like protein 2 OS=Saccharom	-1.43223119	0.31935629	9.80503305	-4.484744	0.00090644	1
P54114	sp|P54114|ALDH3_YEAST Aldehyde dehydrogenase [NAD(P)+] 2	-0.90756698	0.20353216	24.1398147	-4.459084	0.00102203	4
P46956	sp|P46956|PHO86_YEAST Inorganic phosphate transporter PHO	0.95975067	0.22278325	20.1481393	4.308002	0.00163637	3
P31382	sp|P31382|PMT2_YEAST Dolichyl-phosphate-mannose--protein	1.21459467	0.28774793	12.0774591	4.221037	0.0024156	1
P39107	sp|P39107|MNN9_YEAST Mannan polymerase complexes subu	0.95385722	0.23602452	17.9508806	4.041348	0.00527907	2
P34227	sp|P34227|PRX1_YEAST Mitochondrial peroxiredoxin PRX1 OS=	-0.72585541	0.18965659	27.8012372	-3.827209	0.01072839	8
P38281	sp|P38281|APD1_YEAST Actin patches distal protein 1 OS=Sacch	-0.8622131	0.24134855	17.1676408	-3.572481	0.02927145	2
P33315	sp|P33315|TKT2_YEAST Transketolase 2 OS=Saccharomyces cer	-0.94486846	0.27330747	13.3874237	-3.457163	0.04518799	1

SILAC LTQ — **A* vs B**

#acceso	FASTAProteinDescription	Xq-X	SD_Xq	Wq	Zq	FDRq	pep_per_protein
P38143	sp|P38143|GPX2_YEAST Glutathione peroxidase 2 OS=Saccharo	1.793859	0.16326154	37.5173568	10.98764	0	1
Q04120	sp|Q04120|TSA2_YEAST Peroxiredoxin TSA2 OS=Saccharomyces	2.32144492	0.15431333	41.9945699	15.04371	0	2
Q12250	sp|Q12250|RPN5_YEAST 26S proteasome regulatory subunit RP	2.43901717	0.1860154	28.9002903	13.11191	0	1
Q04697	sp|Q04697|GSF2_YEAST Glucose-signaling factor 2 OS=Saccharc	-1.1013179	0.17351874	33.2129323	-6.346968	4.27E-08	1
P33754	sp|P33754|SEC66_YEAST Translocation protein SEC66 OS=Sacch	-1.16005888	0.18458796	29.3489983	-6.284586	6.39E-08	1
P28007	sp|P28007|GAR1_YEAST H/ACA ribonucleoprotein complex sub	-1.03300818	0.1772621	31.8249861	-5.827575	1.09E-06	1
Q07800	sp|Q07800|PSR1_YEAST Phosphatase PSR1 OS=Saccharomyces	-1.08390282	0.22496707	19.7588695	-4.818051	0.00016112	1
Q03558	sp|Q03558|OYE2_YEAST NADPH dehydrogenase 2 OS=Saccharo	-0.66057306	0.13722669	53.1034596	-4.813736	0.00016464	12
P00431	sp|P00431|CCPR_YEAST Cytochrome c peroxidase, mitochondri	0.75714304	0.15968018	39.2191331	4.741622	0.00023564	2
P10127	sp|P10127|ADH4_YEAST Alcohol dehydrogenase 4 OS=Saccharc	-0.64457351	0.14056173	50.6134345	-4.585697	0.00035203	7
P36139	sp|P36139|PET10_YEAST Protein PET10 OS=Saccharomyces cer	0.67199368	0.15241225	43.0487213	4.409053	0.00080775	2
Q3E841	sp|Q3E841|YN034_YEAST Uncharacterized protein YNR034W-A	-0.70984865	0.16683208	35.9286472	-4.254869	0.00125181	2
P04046	sp|P04046|PUR1_YEAST Amidophosphoribosyltransferase OS=S	-0.75873193	0.17880793	31.277098	-4.243279	0.00131827	1
P07991	sp|P07991|OAT_YEAST Ornithine aminotransferase OS=Sacchar	-0.61099335	0.14286204	48.9966436	-4.276807	0.00147503	5
P07262	sp|P07262|DHE4_YEAST NADP-specific glutamate dehydrogena	-0.55148135	0.13805309	52.4696011	-3.994705	0.00314967	10
P31382	sp|P31382|PMT2_YEAST Dolichyl-phosphate-mannose--protein	-0.63888668	0.16219086	38.0143195	-3.939104	0.00334894	2
Q00711	sp|Q00711|DHSA_YEAST Succinate dehydrogenase [ubiquinone	0.71052633	0.17859712	31.3509776	3.978375	0.00337399	1
P53083	sp|P53083|MDM34_YEAST Mitochondrial distribution and morp	-0.83001349	0.20686751	23.3676701	-4.012295	0.00359863	1
P11986	sp|P11986|INO1_YEAST Inositol-3-phosphate synthase OS=Sacc	-0.53565412	0.13563697	54.3555454	-3.949175	0.00381322	25
P22803	sp|P22803|TRX2_YEAST Thioredoxin-2 OS=Saccharomyces cere	0.54499411	0.14271177	49.099881	3.818845	0.00549015	4
P47912	sp|P47912|LCF4_YEAST Long-chain-fatty-acid--CoA ligase 4 OS=	0.59523808	0.16688234	35.9070117	3.566813	0.0147963	1
P29496	sp|P29496|MCM5_YEAST Minichromosome maintenance prote	-0.6943741	0.19729889	25.6892083	-3.519402	0.01529553	1
Q04371	sp|Q04371|YMR7_YEAST UPF0364 protein YMR027W OS=Sacch	-0.56716107	0.16589255	36.3367647	-3.418846	0.02223927	1
P07258	sp|P07258|CARA_YEAST Carbamoyl-phosphate synthase arginir	0.56308499	0.17697558	31.9281157	3.18171	0.04556235	1

SILAC Orbitrap — **A* vs B**

#acceso	FASTAProteinDescription	Xq-X	SD_Xq	Wq	Zq	FDRq	pep_per_protein
P18900	sp|P18900|COQ1_YEAST Hexaprenyl pyrophosphate synthetase	1.66635503	0.16777449	35.5261498	9.932112	0	1
P38143	sp|P38143|GPX2_YEAST Glutathione peroxidase 2 OS=Saccharo	1.71220069	0.15604316	41.0686617	10.97261	0	1
Q02354	sp|Q02354|UTP6_YEAST U3 small nucleolar RNA-associated pro	-3.18142316	0.18000085	30.8639073	-17.67449	0	1
Q04120	sp|Q04120|TSA2_YEAST Peroxiredoxin TSA2 OS=Saccharomyces	2.32262731	0.13129903	58.0065409	17.6896	0	3
Q05930	sp|Q05930|MDM30_YEAST Mitochondrial distribution and mor	2.32987918	0.16793181	35.4596197	13.87396	0	1
Q04951	sp|Q04951|SCW10_YEAST Probable family 17 glucosidase SCW1	1.55914253	0.19344827	26.7220836	8.059739	2.22E-13	1
P40081	sp|P40081|MGDP1_YEAST Putative magnesium-dependent pho	1.61689077	0.2178842	21.0643721	7.420872	2.91E-11	1
P53327	sp|P53327|SLH1_YEAST Antiviral helicase SLH1 OS=Saccharomy	1.16839432	0.16641703	36.1080864	7.020882	5.51E-10	2
P41816	sp|P41816|OYE3_YEAST NADPH dehydrogenase 3 OS=Saccharo	1.28223754	0.21682793	21.2701014	5.913618	5.57E-07	1

P00431	sp|P00431|CCPR_YEAST Cytochrome c peroxidase, mitochondri	0.73813506	0.1250705	63.9278717	5.901752	5.99E-07	4
P10127	sp|P10127|ADH4_YEAST Alcohol dehydrogenase 4 OS=Saccharc	-0.62775055	0.1123309	79.2504139	-5.588405	3.82E-06	11
Q03558	sp|Q03558|OYE2_YEAST NADPH dehydrogenase 2 OS=Saccharo	-0.63907171	0.1158195	74.5481143	-5.517825	4.29E-06	7
P36156	sp|P36156|GTO2_YEAST Glutathione S-transferase omega-like 2	0.69119358	0.13906515	51.7086705	4.970286	8.35E-05	2
P11745	sp|P11745|RNA1_YEAST Ran GTPase-activating protein 1 OS=Sa	0.55466675	0.11344122	77.7066586	4.889464	0.0001263	10
P38009	sp|P38009|PUR92_YEAST Bifunctional purine biosynthesis prote	-0.568238	0.1218816	67.3168329	-4.662213	0.00031262	4
P36006	sp|P36006|MYO3_YEAST Myosin-3 OS=Saccharomyces cerevisia	0.73905074	0.15868312	39.7135364	4.6574	0.00032001	2
P05150	sp|P05150|OTC_YEAST Ornithine carbamoyltransferase OS=Sac	0.56644088	0.12457576	64.4366421	4.546959	0.00045325	4
P18239	sp|P18239|ADT2_YEAST ADP,ATP carrier protein 2 OS=Saccharc	0.52205547	0.115099	75.4843435	4.535708	0.0004781	7
P32337	sp|P32337|IMB3_YEAST Importin subunit beta-3 OS=Saccharom	0.56113685	0.12393882	65.1006432	4.527531	0.00049698	4
Q07896	sp|Q07896|NOC3_YEAST Nucleolar complex-associated protein	0.64084548	0.14050819	50.6520154	4.560912	0.00050898	2
P07991	sp|P07991|OAT_YEAST Ornithine aminotransferase OS=Sacchar	-0.49968697	0.11137384	80.6182965	-4.486574	0.00051664	15
P11986	sp|P11986|INO1_YEAST Inositol-3-phosphate synthase OS=Sacc	-0.47626586	0.10855881	84.8535123	-4.387169	0.00081971	29
P38113	sp|P38113|ADH5_YEAST Alcohol dehydrogenase 5 OS=Saccharc	0.69282139	0.16036886	38.8830122	4.320174	0.00097376	1
Q03529	sp|Q03529|SCS7_YEAST Inositolphosphorylceramide-B C-26 hyc	-0.68124983	0.15834351	39.8840694	-4.302354	0.0010555	1
P07262	sp|P07262|DHE4_YEAST NADP-specific glutamate dehydrogena	-0.47681298	0.11028408	82.2194149	-4.323498	0.00109623	17
Q12512	sp|Q12512|ZPS1_YEAST Protein ZPS1 OS=Saccharomyces cerevi	0.59534349	0.139707	51.2346395	4.261372	0.00117139	2
P29478	sp|P29478|SEC65_YEAST Signal recognition particle subunit SEC	-0.75353705	0.17728648	31.8162327	-4.250392	0.00123031	1
P18544	sp|P18544|ARGD_YEAST Acetylornithine aminotransferase, mit	0.57718429	0.14951593	44.7326924	3.860353	0.00585249	2
P40043	sp|P40043|YEP7_YEAST Uncharacterized protein YER067W OS=	-0.51244009	0.13206562	57.3350838	-3.880193	0.00601753	3
P50278	sp|P50278|SOL1_YEAST Probable 6-phosphogluconolactonase 1	0.71197319	0.18646828	28.7600784	3.8182	0.00694859	1
P49334	sp|P49334|TOM22_YEAST Mitochondrial import receptor subur	0.58465694	0.15633312	40.916457	3.739815	0.00862657	1
P53866	sp|P53866|SQS1_YEAST Protein SQS1 OS=Saccharomyces cerev	-0.72799868	0.19341754	26.7305734	-3.763871	0.00864784	1
Q04175	sp|Q04175|SXM1_YEAST Importin beta SMX1 OS=Saccharomyc	0.63178162	0.16905715	34.9891114	3.737089	0.00872059	1
P38777	sp|P38777|FSH1_YEAST Family of serine hydrolases 1 OS=Sacch	-0.47001629	0.1258372	63.1512425	-3.735114	0.0087893	4
P22768	sp|P22768|ASSY_YEAST Argininosuccinate synthase OS=Sacchar	0.4256425	0.11556383	74.8783294	3.683181	0.00986519	8
Q03104	sp|Q03104|MSC1_YEAST Meiotic sister chromatid recombinatic	0.45077461	0.12338901	65.6820999	3.65328	0.01108882	4
P40579	sp|P40579|YIV5_YEAST Uncharacterized oxidoreductase YIR035	-0.54348804	0.15203113	43.2648229	-3.574847	0.01500857	2
Q06440	sp|Q06440|CORO_YEAST Coronin-like protein OS=Saccharomyc	0.45507843	0.12915237	59.9508399	3.523578	0.01679521	3
P09232	sp|P09232|PRTB_YEAST Cerevisin OS=Saccharomyces cerevisiae	-0.43207957	0.12301975	66.076997	-3.512278	0.01752578	4
P11972	sp|P11972|SST2_YEAST Protein SST2 OS=Saccharomyces cerevis	0.58477756	0.16798175	35.4385385	3.481197	0.01825045	1
P22803	sp|P22803|TRX2_YEAST Thioredoxin-2 OS=Saccharomyces cere	0.44941596	0.12921496	59.8927722	3.478049	0.01846614	3
P53278	sp|P53278|YG3A_YEAST Uncharacterized protein YGR130C OS=	0.53381159	0.15408206	42.1207275	3.464463	0.01942456	3
P47137	sp|P47137|YJ66_YEAST Uncharacterized oxidoreductase YJR096	0.4425831	0.12846902	60.5903091	3.445057	0.01945088	3
Q12207	sp|Q12207|NCE2_YEAST Non-classical export protein 2 OS=Sacc	0.5884114	0.16890163	35.0535746	3.483752	0.01950422	1
P36132	sp|P36132|KAE1_YEAST Putative glycoprotein endopeptidase K	-0.58770095	0.17357145	33.1927648	-3.385931	0.024167	1
P54114	sp|P54114|ALDH3_YEAST Aldehyde dehydrogenase [NAD(P)+] 2	0.40169112	0.11964469	69.8575215	3.357367	0.02509668	5
P06115	sp|P06115|CATT_YEAST Catalase T OS=Saccharomyces cerevisia	0.38118357	0.11305684	78.2359414	3.37161	0.02545922	11
P46367	sp|P46367|ALDH4_YEAST Potassium-activated aldehyde dehydr	-0.38410769	0.11487285	75.7818544	-3.343764	0.02636009	8
P38333	sp|P38333|ENP1_YEAST Essential nuclear protein 1 OS=Sacchar	0.49740092	0.15011748	44.3749063	3.313411	0.02656849	2
P30822	sp|P30822|XPO1_YEAST Exportin-1 OS=Saccharomyces cerevisia	0.41446433	0.12577584	63.2128798	3.295262	0.0267994	4
Q05785	sp|Q05785|ENT2_YEAST Epsin-2 OS=Saccharomyces cerevisiae (	-0.60651585	0.18339089	29.7333941	-3.30723	0.0271618	1
P38774	sp|P38774|DOG1_YEAST 2-deoxyglucose-6-phosphate phospha	-0.4496175	0.13527852	54.6439793	-3.323643	0.02718083	2
P36013	sp|P36013|MAOM_YEAST NAD-dependent malic enzyme, mitoc	0.4417416	0.13303318	56.504111	3.320537	0.02748507	3
P32795	sp|P32795|YME1_YEAST Protein YME1 OS=Saccharomyces cere	0.44523314	0.13480667	55.0271761	3.302753	0.02759912	3
P54000	sp|P54000|SUB1_YEAST RNA polymerase II transcriptional coac	-0.59728318	0.18008672	30.8344798	-3.316642	0.02787107	1
Q03337	sp|Q03337|TRS31_YEAST Transport protein particle 31 kDa sub	-0.80566099	0.24803696	16.2542603	-3.248149	0.03165849	1
P26263	sp|P26263|PDC6_YEAST Pyruvate decarboxylase isozyme 3 OS=	-0.4746901	0.14770961	45.8334432	-3.213671	0.03386957	3
P02557	sp|P02557|TBB_YEAST Tubulin beta chain OS=Saccharomyces c	0.36124132	0.11240163	79.150699	3.213844	0.03569555	12
Q04439	sp|Q04439|MYO5_YEAST Myosin-5 OS=Saccharomyces cerevisia	0.52474534	0.16485096	36.7973908	3.18315	0.03765126	1
Q03264	sp|Q03264|NGL2_YEAST RNA exonuclease NGL2 OS=Saccharom	0.64029476	0.20178979	24.558487	3.173078	0.03898227	1
P47077	sp|P47077|YJB0_YEAST Pumilio domain-containing protein YJL0	0.48077518	0.15497978	41.6341706	3.10218	0.04720645	2

iTRAQ LTQ(PQD) — **A vs B** — **116 vs 115**

# acceso	FASTAProteinDescription	Xq-X	SD_Xq	Wq	Zq	FDRq	pep_per_protein
P29509	sp|P29509|TRXB1_YEAST Thioredoxin reductase 1 OS=Saccharo	-1.09401634	0.167781	35.5233935	-6.520502	7.29E-08	8
Q04120	sp|Q04120|TSA2_YEAST Peroxiredoxin TSA2 OS=Saccharomyces	-1.72671947	0.27839883	12.9022421	-6.202323	5.79E-07	2
P22803	sp|P22803|TRX2_YEAST Thioredoxin-2 OS=Saccharomyces cere	-1.20303858	0.21485989	21.6615378	-5.599177	7.47E-06	3
Q3E7X9	sp|Q3E7X9|RS28A_YEAST 40S ribosomal protein S28-A OS=Sacc	-1.39220726	0.26504204	14.2354261	-5.252779	5.19E-05	2
Q04178	sp|Q04178|HPRT_YEAST Hypoxanthine-guanine phosphoribosy	-1.86586532	0.37530208	7.09966834	-4.971636	0.00013809	1
P09435	sp|P09435|HSP73_YEAST Heat shock protein SSA3 OS=Saccharo	-0.82304624	0.19791316	25.5299894	-4.158623	0.00475684	3
P48837	sp|P48837|NUP57_YEAST Nucleoporin NUP57 OS=Saccharomyc	1.32510136	0.31846684	9.8598784	4.160877	0.0065942	2
P26781	sp|P26781|RS11_YEAST 40S ribosomal protein S11 OS=Saccharc	-0.54726971	0.13740856	52.9629823	-3.982792	0.01011925	9
P39935	sp|P39935|IF4F1_YEAST Eukaryotic initiation factor 4F subunit p	1.42452057	0.37049064	7.28526785	3.844957	0.0113997	2
P38181	sp|P38181|NU170_YEAST Nucleoporin NUP170 OS=Saccharomy	1.33966662	0.34951366	8.18599922	3.832945	0.01197119	1
P53337	sp|P53337|ERV29_YEAST ER-derived vesicles protein ERV29 OS=	-1.39678536	0.36278839	7.59789419	-3.850138	0.01364151	2
Q02803	sp|Q02803|OAZ_YEAST Ornithine decarboxylase antizyme OS=S	1.40368077	0.36462233	7.52165598	3.849684	0.01366681	1
P22217	sp|P22217|TRX1_YEAST Thioredoxin-1 OS=Saccharomyces cere	-0.75502962	0.20566586	23.6415297	-3.671147	0.01931719	3
P07262	sp|P07262|DHE4_YEAST NADP-specific glutamate dehydrogena	0.43782196	0.12512216	63.875091	3.499156	0.03236015	14
O14455	sp|O14455|RL36B_YEAST 60S ribosomal protein L36-B OS=Saccl	-0.72752496	0.20714583	23.3049192	-3.512139	0.03556123	4
P40825	sp|P40825|SYAC_YEAST Alanyl-tRNA synthetase, cytoplasmic OS	0.4416142	0.13019064	58.9984336	3.392058	0.04243796	12
Q02753	sp|Q02753|RL21A_YEAST 60S ribosomal protein L21-A OS=Sacc	-0.53958918	0.15919517	39.4584665	-3.389482	0.04283869	5
P53128	sp|P53128|MTHR2_YEAST Methylenetetrahydrofolate reductas	1.11675626	0.33267288	9.03577057	3.35692	0.04314129	1
Q00955	sp|Q00955|ACAC_YEAST Acetyl-CoA carboxylase OS=Saccharom	0.46479499	0.13628917	53.8365629	3.410359	0.04498172	15

iTRAQ LTQ(PQD) — **B vs A** — **117 - 116**

#acceso	FASTAProteinDescription	Xq-X	SD_Xq	Wq	Zq	FDRq	pep_per_protein
P48837	sp|P48837|NUP57_YEAST Nucleoporin NUP57 OS=Saccharomyc	-2.46959277	0.38746034	6.66109291	-6.373795	1.88E-07	2
Q3E7X9	sp|Q3E7X9|RS28A_YEAST 40S ribosomal protein S28-A OS=Sacc	1.97266406	0.32031619	9.7463548	6.15849	7.50E-07	2
Q04120	sp|Q04120|TSA2_YEAST Peroxiredoxin TSA2 OS=Saccharomyces	1.74481557	0.33394449	8.96708783	5.224867	0.00017794	2
P39935	sp|P39935|IF4F1_YEAST Eukaryotic initiation factor 4F subunit p	-2.13421755	0.43195509	5.35948178	-4.940832	0.00019856	2
P09435	sp|P09435|HSP73_YEAST Heat shock protein SSA3 OS=Saccharo	1.12531331	0.24294712	16.9424611	4.631927	0.00092471	3
P40482	sp|P40482|SEC24_YEAST Protein transport protein SEC24 OS=Sa	-1.31453285	0.30504195	10.7468417	-4.309351	0.00278621	3
P05747	sp|P05747|RL29_YEAST 60S ribosomal protein L29 OS=Saccharc	1.32744142	0.31254051	10.2373459	4.247262	0.00368239	2
P29509	sp|P29509|TRXB1_YEAST Thioredoxin reductase 1 OS=Saccharo	0.82383226	0.20060619	24.8491383	4.106714	0.00512196	7
P22217	sp|P22217|TRX1_YEAST Thioredoxin-1 OS=Saccharomyces cere	0.97331322	0.25059183	15.9245134	3.884058	0.01311056	3

P26781	sp|P26781|RS11_YEAST 40S ribosomal protein S11 OS=Saccharc	0.63635858	0.16624691	36.182024	3.827792	0.01320133	9
P43616	sp|P43616|CPGL_YEAST Glutamate carboxypeptidase-like prote	-0.78027435	0.20402066	24.0243539	-3.824487	0.01337968	5
P40825	sp|P40825|SYAC_YEAST Alanyl-tRNA synthetase, cytoplasmic O	-0.57596003	0.15266476	42.9064311	-3.772711	0.01373956	12
Q08745	sp|Q08745|RS10A_YEAST 40S ribosomal protein S10-A OS=Sacc	0.89375384	0.24418123	16.7716368	3.660207	0.02144202	4
Q00955	sp|Q00955|ACAC_YEAST Acetyl-CoA carboxylase OS=Saccharom	-0.57813312	0.16133418	38.4191052	-3.583451	0.02472897	15
Q05567	sp|Q05567|SGPL_YEAST Sphingosine-1-phosphate lyase OS=Sac	1.23348833	0.35704417	7.84433475	3.454722	0.03515183	2
P05756	sp|P05756|RS13_YEAST 40S ribosomal protein S13 OS=Saccharc	0.68391377	0.19661548	25.8681031	3.478433	0.03678185	5
P05759	sp|P05759|RS37_YEAST 40S ribosomal protein S31 OS=Saccharc	0.71141268	0.21120672	22.4173625	3.368324	0.04289721	4
P61864	sp|P61864|UBIQ_YEAST Ubiquitin OS=Sacc	0.90151002	0.26735736	13.9899359	3.371929	0.04763199	3

Documentación auto-generada de ejemplo del esquema QuiXML

XML Schema Documentation

Table of Contents

top

Schema Document Properties

Target Namespace	None

Element and Attribute Namespaces	• Global element and attribute declarations belong to this schema's target namespace. • By default, local element declarations belong to this schema's target namespace. • By default, local attribute declarations have no namespace.

Declared Namespaces

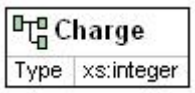
Prefix	Namespace
xml	http://www.w3.org/XML/1998/namespace
xs	http://www.w3.org/2001/XMLSchema

+ Schema Component Representation

top

Global Declarations

Element: Charge

Name	Charge
Type	xs:integer
Nillable	no
Abstract	no
Diagram	

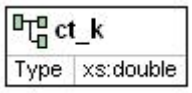
- XML Instance Representation

```
<Charge> xs:integer </Charge>
```

+ Schema Component Representation

top

Element: ct_k

Name	ct_k
Type	xs:double
Nillable	no
Abstract	no
Diagram	

- XML Instance Representation

```
<ct_k> xs:double </ct_k>
```