

# Proteómica: hacia el entendimiento del lenguaje de las proteínas

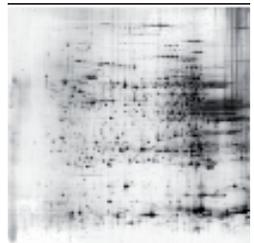
Victoria Pando Robles y César Ferreira Batista

La revolución en biología molecular de los años 80, ejemplificada por la facilidad con la que actualmente se realizan amplificaciones, clonaciones y secuenciación automatizada de ADN, ha permitido en corto tiempo conocer el genoma completo de distintos organismos, incluido el del ser humano (<http://www.ncbi.nlm.nih.gov/>).

Actualmente disponemos de un listado total de genes que definen el genoma de una especie, conocemos la secuencia de nucleótidos de sus genes y, a partir de ellas, podemos inferir con cierta confiabilidad las proteínas que se expresan a partir de ese genoma. Sin embargo, este conocimiento no nos permite entender la función biológica del gen. Por ejemplo: ¿por qué el mono y el hombre son tan diferentes, si comparten el 99% de su genoma? o ¿qué condiciones fisiológicas cambian en una célula en un cuadro de enfermedad? En resumen, del análisis de los proyectos genómicos se ha aprendido que el número de genes contenido en el genoma de un organismo no correlaciona con la complejidad del mismo y se infiere que la complejidad morfológica y funcional de eucariotes superiores como el ser humano depende de la regulación de la expresión genética y de las interacciones entre sus proteínas.

Las proteínas son moléculas orgánicas complejas, formadas por aminoácidos ordenados en largas hileras o cadenas polipeptídicas mantenidas por enlaces químicos entre el grupo amino ( $\text{NH}_2$ ) de un aminoácido y el grupo carboxilo ( $\text{COOH}$ ) del siguiente aminoácido. A este tipo de enlace se denomina enlace peptídico. Se conoce que en la naturaleza existen 20 aminoácidos, las posibilidades de ordenamiento de estos aminoácidos en las proteínas es  $20^{20}$  ( $1 \times 10^{26}$ ), es decir que la maquinaria celular puede producir millones de péptidos diferentes con sólo 20 aminoácidos, mientras que los 4 nucleótidos que conforman el ADN se ordenan en sólo 256 posibilidades. Estos números evidencian la complejidad y variabilidad estructural que pueden presentar las proteínas.

Las proteínas se encuentran en todos los organismos, son las moléculas biológicas más abundantes; por ejemplo, en la bacteria *E. coli* constituyen más del 50% de su peso seco, mientras que otras moléculas como el ADN y el ARN constituyen el 3% y el 20% respectivamente. Al igual que otras macromoléculas, como los ácidos nucleicos y los polisacáridos (azúcares), las proteínas son esenciales para el funcionamiento celular y, por ende, para la vida. Debido a su heterogeneidad estructural, las proteínas par-



ticipan en diferentes procesos celulares; muchas son enzimas que catalizan diferentes reacciones químicas vitales para el metabolismo celular, otras tienen un papel estructural, como las proteínas del citoesqueleto, que mantienen la estructura y forma celular. Las proteínas son fundamentales en la comunicación celular, vía transducción de señales; en la respuesta inmune; en el mantenimiento de la homeostasis celular y en el ciclo celular, ya que controlan y realizan la replicación de las moléculas de ADN permitiendo así que la información se conserve de una generación a otra. Las proteínas son también un componente necesario de la dieta, ya que los animales no pueden sintetizar todos sus aminoácidos. Sin embargo, otros organismos, como las plantas, pueden fabricar todos los aminoácidos por medio de la fotosíntesis.

Las proteínas son fabricadas en la célula a partir de la información genética codificada en el ADN, de acuerdo con el dogma central de la biología molecular. Cada proteína tiene su propia y única secuencia de aminoácidos, que es especificada por la secuencia de nucleótidos del gen que la codifica. Los genes contenidos en el ADN son primero transcritos en ARN por una enzima denominada ARN polimerasa. La información contenida en el ARN es interpretada por una molécula transportadora de aminoácidos, conocida como ARN de transferencia, y los ribosomas son los encargados de formar la cadena polipeptídica que constituye la proteína. Las proteínas son sintetizadas del extremo amino al extremo carboxilo. El mensaje contenido en el ARN es leído por la maquinaria de síntesis proteínica mediante un código genético universal que se lee en tripletes o codones. Existen 64 codones que codifican para los 20 aminoácidos, por lo que algunos aminoácidos son codificados por más de un codón. Por ejemplo: los codones CCU, CUC, CUA, CUG, UUA y UUG codifican para el aminoácido leucina. El código genético especifica la síntesis de 20 aminoácidos, sin embargo, estos residuos pueden ser alterados químicamente en la proteína, a este evento se conoce como modificación postraduccional; éstas se reali-

zan en diferentes compartimentos celulares, como citoplasma, núcleo, retículo endoplásmico y aparato de Golgi. Las modificaciones postraduccionales son utilizadas en la célula como un mecanismo de control, ya que éstas señalan la localización de las proteínas y pueden alterar su función. En los procariotes, como las bacterias, el ARN se sintetiza en el citoplasma y puede ser inmediatamente traducido en proteínas, mientras que en eucariotes, el ARN se sintetiza en el núcleo y se transloca a través de la membrana nuclear en el citoplasma, lugar donde se realiza la síntesis de proteínas. La velocidad de síntesis de proteínas es más rápida en bacterias que en animales superiores; éstas pueden ensamblar 20 aminoácidos por segundo. El tamaño de las proteínas puede ser medido por el número de aminoácidos que contienen o por su masa total, la cual es normalmente reportada en unidades de masa atómica dalton. Un dalton (da) equivale a 1/12 de la masa atómica del  $^{12}\text{C}$  (carbono doce). Por ejemplo: las proteínas de la levadura tienen en promedio 466 a-a de tamaño y 53 Kda de masa. La proteína de mayor peso molecular conocida hasta ahora es la titina, componente del sarcómero muscular, tiene una masa de 3000 Kda y está formada por una cadena de 27 000 aminoácidos.

### Proteómica

El término proteoma fue acuñado en 1994 para definir a todas las proteínas que son expresadas por un genoma en un tejido o en una célula. El proteoma de un organismo es un elemento altamente dinámico, ya que sus componentes varían dependiendo del tejido, célula o compartimento celular que se estudie y estos, a su vez, pueden cambiar debido a alteraciones en su ambiente, como situaciones de estrés, acción de fármacos, requerimientos energéticos, o su estado fisiológico (normal o patológico). El crecimiento en el número de proyectos de investigación orientados al estudio de los genomas de forma sistemática, ha dado lugar a la aparición de nuevas tecnologías a gran escala (tipo "high throughput") que, en el caso de las proteínas,

se denomina *Proteómica*. Esta puede dividirse en *proteómica de expresión*, que tiene como objetivo la descripción del proteoma total de un tejido, fluido, tipo celular u organelo y las mediciones cuantitativas de los niveles de expresión proteínica, y *proteómica funcional*, que se encarga del estudio de la función de proteínas dentro de sistemas biológicos (relaciona cambios de expresión con una función determinada) y la regulación de su expresión, incluyendo las interacciones proteínas-proteína, proteínas-ADN, proteínas-ARN y las modificaciones post-traduccionales.

La diferencia principal entre genómica y proteómica, es que el genoma puede ser visto como una colección de genes cuya naturaleza es estática; es decir, no cambia entre célula y célula, mientras que el proteoma es una entidad cambiante, es una colección dinámica de proteínas que difieren de un individuo a otro, o de una célula a otra. Por ejemplo, se predice que el número de proteínas expresadas en el ser humano va de 50 mil a 500 mil. ¿Cómo puede ser eso si el número de genes presentes en nuestro genoma es de 25 mil? En primer lugar, esto se debe a que los genes no necesariamente expresan una sola proteína, por procesamiento diferencial de los transcritos de ARNm (maduración y empalme alternativo) se pueden generar diversas proteínas. Segundo, las proteínas presentan alrededor de 300 diferentes tipos de modificaciones postraduccionales, incluyendo fosforilación, glicosilación, acetilación, deaminación, miristolación, entre otras ([www.abrf.org/index.cfm/dm.home](http://www.abrf.org/index.cfm/dm.home)). Estas modificaciones pueden afectar la estructura, localización, función y recambio, e implican funciones reguladas por factores internos y externos a las células, desencadenando procesos de expresión genética diferencial. Esta consiste en que, pese a que todas las células de un individuo contienen en su interior la misma información genética, se producen diferencias en la intensidad, momento de la expresión, o en la información expresada. Como resultado de estas variaciones se produce la diferenciación celular y pueden originarse determinadas patologías. Es así

que el proteoma humano está definido por muchos proteomas que son característicos de tipos celulares específicos (como las células del hígado) o que representan un estado fisiológico particular, normal o patológico (el proteoma de una persona con cáncer gástrico, es distinto al proteoma de un paciente con anemia, o al de una persona sana).

Para que un niño empiece a leer, primero debe aprender las letras, luego entender que el orden de esas letras forma palabras y que las palabras expresan ideas. Del mismo modo, los investigadores primero descubrieron los nucleótidos, luego que el orden de esos nucleótidos forma un gen y que los genes codifican proteínas. Sin embargo, en la práctica primero se pudo secuenciar a las proteínas. En 1950, el investigador sueco E. Edman desarrolló un método que permite saber el orden de los aminoácidos en una proteína, a la que se conoce como *degradación de Edman*. Durante muchos años esta técnica fue de vital importancia en la investigación bioquímica. La secuenciación parcial de fragmentos proteínicos y el ensamblaje de estas secuencia permitió conocer en forma completa la secuencia de aminoácidos de algunas proteínas. En 1958 F. Sanger obtuvo el Premio Nobel de Química debido a la secuenciación completa de la insulina, proteína involucrada en el metabolismo de la glucosa. Las secuencias parciales obtenidas por Edman también permitieron conocer los extremos amino terminal de muchas proteínas y a, partir de esta secuencia de aminoácidos, construir oligos de ADN que, mediante la acción de la ADN polimerasa, permiten el aislamiento del gene que codifica la proteína. Sin embargo, este método de secuenciación no es compatible con los análisis a gran escala que dominan el campo de las ciencias genómicas. Principalmente porque el número de aminoácidos secuenciados es limitado, en general no se excede de 30-40 a-a secuenciados por muestra, es lento (un ciclo/hora), lo que significa que para secuenciar 30 a-a el equipo se demora 30 horas; ¡imagínense entonces secuenciar una proteína de 1000 aminoácidos, o secuenciar y conocer el proteoma de una célula!

La posibilidad de identificar proteínas en forma global surge gracias a la modernización de la espectrometría de masa (EM). En 1985, J. Fenn y K. Tanaka desarrollaron los sistemas de ionización de macromoléculas ESI y MALDI respectivamente, por lo que se les galardonó con el Premio Nóbel de Química 2002. El método de ionización por electropulverización, ESI (*electrospray ionization*), permite la ionización de moléculas a partir de flujos líquidos bajo aplicación de alto voltaje, mientras que la ionización por el método MALDI, produce iones por bombardeo con rayos láser de muestras en estado sólido auxiliado por matrices cristalizables. Estas metodologías solucionaron la dificultad de generar iones a partir de analitos no volátiles, como las proteínas y polímeros. Desde entonces, la EM desplazó a la secuenciación de proteínas por Edman, debido a su sensibilidad (pmol-fmol), exactitud (100-5 ppm) y rapidez (minutos-segundos); esta varía de acuerdo a la complejidad del análisis y al equipo que se utilice. La espectrometría de masas permite la identificación de proteínas, el conocimiento de su estructura primaria, es decir, la secuencia de los aminoácidos que la conforman, la identificación de sus modificaciones post-traduccionales, y la cuantificación de la expresión proteínica (proteómica cuantitativa). Por otro lado, los resultados obtenidos de la identificación de proteínas de un genoma determinado han permitido corregir datos genómicos. Los genes se predicen en un genoma mediante el uso de técnicas de bioinformática basadas en características comunes de miles de genes analizados. Sin embargo, estas predicciones no siempre son ciertas, sobre todo en genes pequeños o en genes que no tienen homólogos en otros genomas; por ejemplo, la tasa de error en la anotación de 340 genes del genoma de *Mycoplasma genitalum* es del 8%, si extrapolamos ese error en el genoma humano, imagínense las consecuencias. Por otro lado, es importante resaltar que la espectrometría de masas es una técnica analítica con múltiples usos; realiza la medición de masas

de diferentes moléculas, permitiendo así la identificación de nucleótidos, carbohidratos, lípidos y polímeros sintéticos.

### Herramientas de la proteómica

Desafortunadamente, las mismas características que otorgan a las proteínas su papel fundamental como moléculas efectoras de la función celular (diversidad química y estructural y abundancia relativa) también dificultan su análisis experimental. Actualmente no existe un diagrama de flujo único para el análisis proteómico de una muestra, ya que las variables como complejidad, método de separación, concentración y estabilidad de las proteínas, además de la plataforma tecnológica disponible para su análisis, y muy especialmente el tipo de pregunta biológica que se pretende abordar, son los parámetros básicos que determinan la elección de una estrategia de estudio. Por lo tanto, no existe una metodología idónea para el estudio de proteomas en forma sistemática. En consecuencia, la investigación proteómica es el resultado de la aplicación de un conjunto de técnicas que permiten el estudio de proteínas. A continuación, se describirá en forma resumida una metodología para el análisis proteómico de una muestra (**figura 1**).

1) Separación de las proteínas. Una muestra proveniente de un sistema biológico es una mezcla compleja de proteínas. Por lo tanto, las muestras provenientes de células, tejidos u otro tipo de muestras biológicas (sangre, orina, leche, líquido céfalo-raquídeo, semen, saliva, lágrimas, etc.) se separan principalmente por técnicas cromatográficas y/o electroforéticas, las cuales son tecnologías robustas, versátiles y con alta capacidad de resolución. Las más utilizadas son: electroforesis unidimensional (SDS-PAGE) y bidimensional (2-D PAGE), electroforesis capilar, cromatografía líquida de alta resolución (HPLC), cromatografía de afinidad, cromatografía de exclusión molecular y cromatografía de intercambio iónico. La electroforesis bidimensional, por ejemplo, permite la separación de hasta 2000 proteínas en un solo gel,

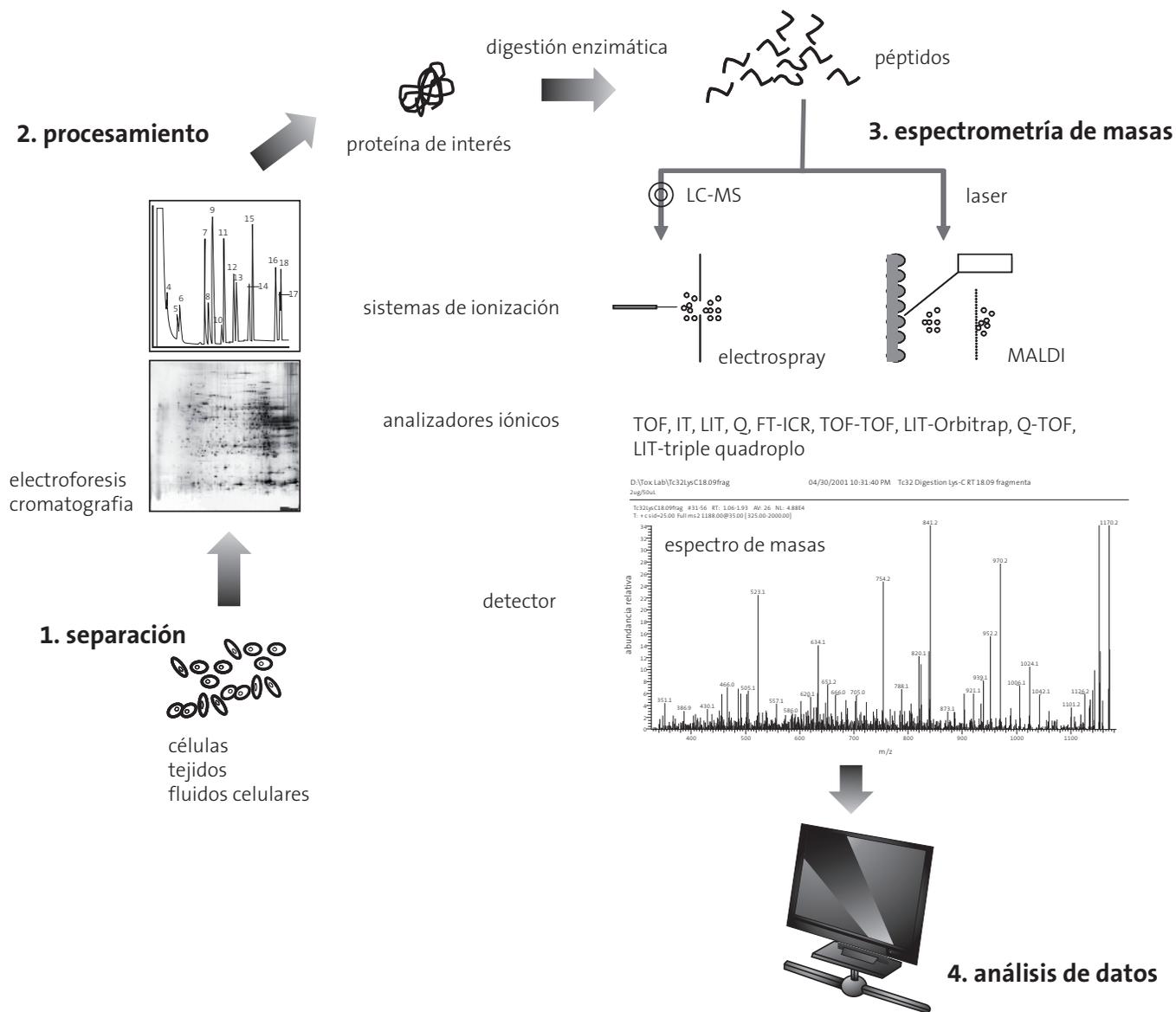


Figura 1. Metodología simplificada del análisis proteómico.

mientras que la electroforesis unidimensional solamente 100. Estas técnicas de aislamiento pueden utilizarse separadamente, en conjunto o en sistemas de flujo continuo, como la cromatografía multidimensional, también conocida como MudPit (*multidimensional protein isolation technology*), que contiene fase reversa e intercambio iónico en forma conjugada. La tecnología de cromatografía líquida acoplada a los espectrómetros de masas (LC-MS, *liquid chromatography-mass spectrometry*) es utilizada para la separación de péptidos provenientes de proteólisis enzimática, péptidos sintéticos o péptidos nativos, por columnas de fase reversa micro-capilares acoplado a sistemas de ionización tipo electrospray-MS. Esta tecnología permite la separación de mezclas complejas de péptidos y su análisis simultáneo por espectrometría de masas. Este sistema es compatible con columnas capilares tipo RP<sub>n</sub> y MudPit de nanoflujo, que funcionan al mismo tiempo como agujas de producción de spray bajo la aplicación de alto voltaje. En la actualidad, el sistema LC-MS es una de las más versátiles y poderosas herramientas del análisis proteómico basado en la espectrometría de masas.

2) Procesamiento de la muestra. El primer paso es la reducción de los puentes de disulfuro y la alquilación de las cisteínas. Tiene como objetivo desnaturalizar las proteínas permitiendo que se expongan los sitios específicos de corte enzimático y, a la vez, se evita la formación de dímeros. Estas modificaciones químicas son fundamentales, ya que para el análisis por EM se requiere que las proteínas sean digeridas a péptidos con masas moleculares menores que 3 KDa. Eso se debe a que las tecnologías de fragmentación de proteínas en espectrómetros de masas (para obtención de secuencias) aún son limitadas y péptidos con mayores masas moleculares producen informaciones estructurales no interpretables. La muestra, reducida y alquilada, generalmente se digiere con tripsina (enzima que corta específicamente en los C-terminales de lisina y arginina, excepto cuando éstas son seguidas de prolina); la elección de esta endoproteasa como enzima

no es casual. En su mayoría, los cortes trípticos generan péptidos con masas moleculares entre 1 y 2 KDa; permiten la diferenciación entre los amino ácidos lisina y glutamina (128 Da), ya que todas las lisinas estarán posicionadas en el C-terminal de los péptidos trípticos, y propicia la presencia de por lo menos dos cargas positivas en el péptido (N-terminal y C-terminal con K o R) que posibilitarán la transferencia de carga a través de las ligaciones amídicas facilitando la fragmentación. La descripción del proceso de reducción, alquilación y digestión enzimática de una proteína para su análisis por EM es apenas uno de los tantos procedimientos que pueden efectuarse en el análisis proteómico. La identificación de biomarcadores, determinación de modificaciones postraduccionales, cuantificación, identificación de mezclas complejas, entre otras aplicaciones, pueden involucrar decenas de metodologías diferentes para el procesamiento de las muestras.

3) Análisis por espectrometría de masas. Este procedimiento involucra la ionización de los componentes de la muestra en fase gaseosa, la separación de las especies iónicas resultantes de acuerdo a la relación de su masa con su carga eléctrica ( $m/z$ ), utilizando campos electromagnéticos en el vacío. Los espectrómetros de masas poseen tres componentes básicos: un sistema de ionización, un analizador de masas y un detector de iones (**figura 2**). Hasta hace poco tiempo, solamente sustancias de bajo peso molecular y relativamente volatilizables podían ser sometidas a una ionización en fase gaseosa, mientras que especies moleculares de mayores tamaños, tales como las proteínas, no se mantenían disueltas a nivel molecular en fase gaseosa. Sin embargo, el desarrollo de sistemas de ionización suaves, como ESI o MALDI, permitieron el análisis de macromoléculas, por lo que hoy en día la EM es la herramienta principal de la investigación proteómica. Para fines prácticos, la diferencia fundamental de los métodos de ionización es que el sistema MALDI utiliza muestras disueltas en matrices sólidas, mientras que el sistema ESI utiliza muestras en fase líquida para la

generación de iones. Los analizadores de masas tienen múltiples funciones que varían de acuerdo a su tecnología; fundamentalmente se refieren al control de los campos electromagnéticos aplicados, que involucra la separación de iones, la resolución de cargas a nivel isotópico, la fragmentación y la capacidad de operación en polaridades diferentes. Los analizadores de masas más usados son: tiempo de vuelo o TOF (*time of flight*), trampa de iones tridimensional o trampa de Paul o IT (*ion trap*), trampa de iones lineal o LIT (*linear ion trap*), quadropolo (Q), triplequadropolo y FT-ICR (*Fourier transform ion cyclotron resonance*). Actualmente, debido al rápido desarrollo de la tecnología en el campo de la EM y de la proteómica, existen espectrómetros de masas que tienen más de un analizador de iones; éstos se denominan espectrómetros híbridos, como TOF-TOF, LIT-Orbitrap, LIT-ETD, LIT-FT-ICR, Q-TOF, LIT-triple quadropolo, etc. Estos equipos presentan mejor resolución, exactitud, sensibilidad y versatilidad en el análisis de péptidos y proteínas. Por ello, son utilizados para secuenciar y cuantificar proteínas, identificar modificaciones post-traduccionales y, en general, en el estudio de muestras biológicas complejas. Los detectores tienen como función detectar el flujo iónico liberado por el analizador, amplificarlo y transmitir esta señal a la computadora, donde se registra en forma de un espectro de masas, los valores  $m/z$  que indican la masa de los iones son dibujados en el eje de las coordenadas mientras que la intensidad de los mismos en el eje de las abscisas. El espectro de masas evidencia el número de componentes en la muestra y el peso molecular de cada componente.

4) Análisis de datos. La interpretación de los espectros de masas constituye la parte más difícil de esta metodología. En esta etapa, la experiencia y la capacidad teórica de dilucidación estructural del espectrometrista son fundamentales, ya que involucra la interdependencia del conocimiento, de los procesos químicos y bioquímicos aplicados a las muestras de proteínas, con el conocimiento de bioinformática para la búsqueda e identificación de proteínas

en bancos de datos públicos y privados, principalmente porque éste se fundamenta en la correlación de valores y diferencias aritméticas de unidades de masas atómicas con la estructura molecular de las proteínas. Por lo tanto, pese a la formidable e indispensable capacidad de procesamiento de datos que la informática nos ofrece, el chequeo manual e integración de los resultados obtenidos es una práctica muy recomendable.

La identificación de proteínas a través de mapas peptídicos es la metodología más utilizada en los análisis proteómicos de los últimos años, principalmente debido a que la secuencia de una proteína es única y al ser cortada con una endoproteasa origina un patrón peptídico específico. Los valores exactos del conjunto de las masas moleculares de los péptidos es la representación única de una determinada proteína, ya que esta lista de masas moleculares está directamente relacionada con su estructura primaria. El espectro de masas evidencia la relación masa/carga ( $m/z$ ) de los péptidos analizados y se convierte en una huella digital de la proteína analizada (*mass finger printing*). Esta lista de masas moleculares medidas por espectrometría de masas, es sometida a los programas de análisis proteómico (PROSPECTOR, MASCOT, ALDENITE, etc.) para su comparación con las diferentes proteínas de las bases de datos, las cuáles son digeridas *in silico*, generando una lista de masas teóricas. La comparación entre el listado de masas experimentales y teóricas es calificada por algoritmos específicos y a cada comparación se le asigna una calificación. Dependiendo de la exactitud y resolución de cada instrumento, sólo unos cuantos péptidos son necesarios para la identificación de la proteína en las bases de datos. Esta metodología tiene el inconveniente que las proteínas sometidas para su identificación deben provenir de genomas secuenciados, debido a que ésta se realiza por comparación de masas moleculares donde uno o más cambios de aminoácidos en su secuencia pueden originar resultados negativos. Otra manera de identificar proteínas a través de espectrometría de

masas es mediante el análisis de los espectros de fragmentación de uno o más péptidos, los cuales fueron generados por el corte de la proteína con una endoproteasa. Con esta técnica se obtienen datos de secuencia, lo que implica una identificación inequívoca de la proteína en cuestión. Como mencionamos anteriormente, una de las características tecnológicas notables de los espectrómetros de masas es la capacidad de fragmentar ordenadamente los péptidos en diferentes posiciones del enlace peptídico. Con esta técnica se obtiene información estructural de la proteína que se está analizando ya que los aminoácidos que conforman la proteína son secuenciados.

La fragmentación de péptidos se realiza por diferentes metodologías tipo CID (*collision induced dissociation*), ETD (*electron transfer dissociation*), PSD (*post source decay*), entre otras. En la tipo CID se selecciona un ión específico y se hace pasar por una celda de colisión que contiene un gas inerte (argón, helio), el choque con este gas induce la fragmentación. En la técnica ETD, a los iones seleccionados se adicionan electrones a través de un donador, generalmente antraceno cargado, para producir la fragmentación. El PSD utiliza la fragmentación de iones metaestables durante el tiempo de vuelo en TOFs con reflectron. Los iones hijos generados por colisión son separados con base a su relación masa/carga ( $m/z$ ) y registrados. La fragmentación de péptidos, según la metodología utilizada, origina un corte específico del enlace peptídico; consecuentemente, las diferencias de masas moleculares entre los iones generados es la representación de cada aminoácido de la cadena polipeptídica, proporcionando de esta forma la secuencia del péptido (**figura 2**). Mediante la fragmentación de péptidos se determina también la presencia de modificaciones postraduccionales. La factibilidad de este análisis depende del tipo de analizador de iones que posea el espectrómetro de masas.

### Lenguaje proteínico y funcionamiento celular

El funcionamiento celular depende de un complejo juego de interacciones entre cientos de diferentes biomoléculas (ADN, ARN, proteínas, metabolitos, etc.), las cuales se encargan de mantener la integridad, morfología y funcionamiento. Como mencionamos anteriormente, los genes almacenan la información para las funciones básicas de las proteínas; sin embargo, la dinámica celular de las proteínas en tiempo real, y la regulación de la estructura proteínica y su función, se realiza debido a la reversibilidad de las modificaciones postraduccionales (MPTs) específicas presentes en estas moléculas. Durante los últimos años las MPTs han llamado la atención de la comunidad científica en las áreas de la biología y ciencias biomédicas, ya que se ha descubierto una serie de MPTs que antes no se conocían y que presentan funciones vitales para la célula; por ejemplo, la *sumoilación*, modificación relacionada con la regulación transcripcional y el ciclo celular. Además, se ha visto que están presentes en todas las especies, desde las bacterias hasta el humano. Antes se consideraba que eran un mecanismo de regulación en organismos superiores principalmente. En general el estudio de las MPTs se ha visto de manera reduccionista, con la idea de que un sitio modificado representa una función reguladora. Por ello y por las limitaciones tecnológicas, cada tipo de modificación ha sido estudiada en una determinada proteína, en un aminoácido específico o en un número limitado de aminoácidos. Sin embargo, con el advenimiento de las ciencias genómicas y las técnicas de estudio a gran escala, ahora es evidente que las modificaciones químicas en una determinada proteína no sólo pueden mediar una función biológica, sino también funcionan para regular otras modificaciones que originan una función diferente en la misma proteína, o actuar en una sinfonía de interacciones moleculares para modular la estabilidad e integridad de una población de proteínas. Actualmente, la presencia de múltiples sitios de fosforilación se reconoce como

un importante mecanismo de regulación de la localización de las proteínas, de su actividad funcional y de la atenuación de las interacciones proteína-proteína y proteína-ligando; aunque los detalles moleculares de este mecanismo de regulación siguen siendo objeto de estudio. Uno de los aportes más significativos de la aplicación de espectrometría de masas en la biología, ha sido la revelación del “código de las histonas”, el cual regula la estructura y función de los nucleosomas en la cromatina. Este código es un complejo múltiple de sitios de modificación, donde la histona puede presentar reacciones de metilación, acetilación, fosforilación y ubiquitinación. Estas modificaciones modulan la interacción de las histonas con los ácidos nucleicos y con factores proteínicos de regulación durante la replicación, transcripción y reparación del ADN.

Un punto fundamental es el conocimiento de las modificaciones que ocurren en una proteína en un determinado tiempo, bajo determinadas condiciones y cómo estas modificaciones señalizan sus interacciones con otras proteínas. La marca de las MPTs se constituye en un código molecular que dicta la conformación de proteínas, su localización molecular, interacciones macromoleculares y actividades, dependiendo del tipo de célula, tejido o condiciones ambientales. Hoy en día, la metodología idónea para la identificación de MPTs es la espectrometría de masas. Sus resultados permiten dilucidar los principios moleculares y funcionales de la plasticidad de las proteínas, debido a la caracterización y cuantificación de las MPTs. Existen diferentes técnicas para estudiar las MPTs, tantas como la existencia de modificaciones (recuerden que se ha reportado hasta la fecha alrededor de 300 MPTs diferentes). En general, primero se separa y enriquece la población de proteínas que se quiere identificar, para posteriormente realizar su caracterización molecular por espectrometría de masas. Por ejemplo, en el caso de aminoácidos fosforilados, se utilizan para su enriquecimiento columnas IMAC (Fe, Ga,  $\text{TiO}_2$ , etc), columnas de afinidad con anticuerpos espe-

cíficos o soportes sólidos fosfoamidados para el enriquecimiento de fosfopéptidos. Para el aislamiento de péptidos glicosilados, se utilizan lectinas o se generan de sitios específicos de N-glicosilaciones por tratamiento con PN-Gasas y agua pesada ( $\text{H}_2\text{O}^{18}$ ), entre otros. Posteriormente se define el sitio de modificación por espectrometría de masas.

Otro punto importante para dilucidar el lenguaje proteínico es la cuantificación de proteínas presentes en un sistema biológico determinado en un momento particular. Esto se debe, principalmente, a que diferentes estudios demuestran que la regulación celular no está dada por la presencia o ausencia de una o varias proteínas, sino que puede ser regulada por aumentos o disminuciones en la concentración de una población determinada de proteínas. Por ejemplo: la transducción de señales intracelulares mediada por receptores de cinasas de tirosina, tales como el factor de crecimiento epidermal (EGF), fue estudiada por proteómica cuantitativa en células HELA estimuladas con EGF, para identificar y cuantificar los fosfopéptidos que cambian su patrón de expresión en respuesta a la estimulación. Se encontraron 6600 sitios de fosforilación en tirosina en 2244 proteínas; el 40% de los sitios de fosforilación son modulados al menos dos veces después del estímulo con EGF y, sorprendentemente, la mayoría de proteínas que tiene múltiples sitios de fosforilación muestran diferente cinética, lo que sugiere que ellas sirven de plataforma para integrar diferentes señales. Actualmente, la única metodología que permite realizar cuantificaciones a nivel molecular es la proteómica cuantitativa, que básicamente se realiza por dos procedimientos: 1) el marcaje de proteínas con fluoróforos Cy, seguido de la separación y análisis de geles con el sistema DIGE (*difference gel electrophoresis*), finalizando con la identificación de proteínas por EM, y 2) mediante la incorporación de isótopos estables en cultivos celulares (marcaje metabólico) o reactivos químicos con diferentes isótopos que marcan las proteínas para su

posterior análisis por espectrometría de masas. Estas técnicas son: ICAT (*isotope-coded affinity tags*), SILAC (*stable isotope labeling with amino acids in cell cultura*), iTRAQ (*isobaric tags for relative and absolute quantification*) y AQUA (*absolute quantification*), las cuales citamos a modo de referencia.

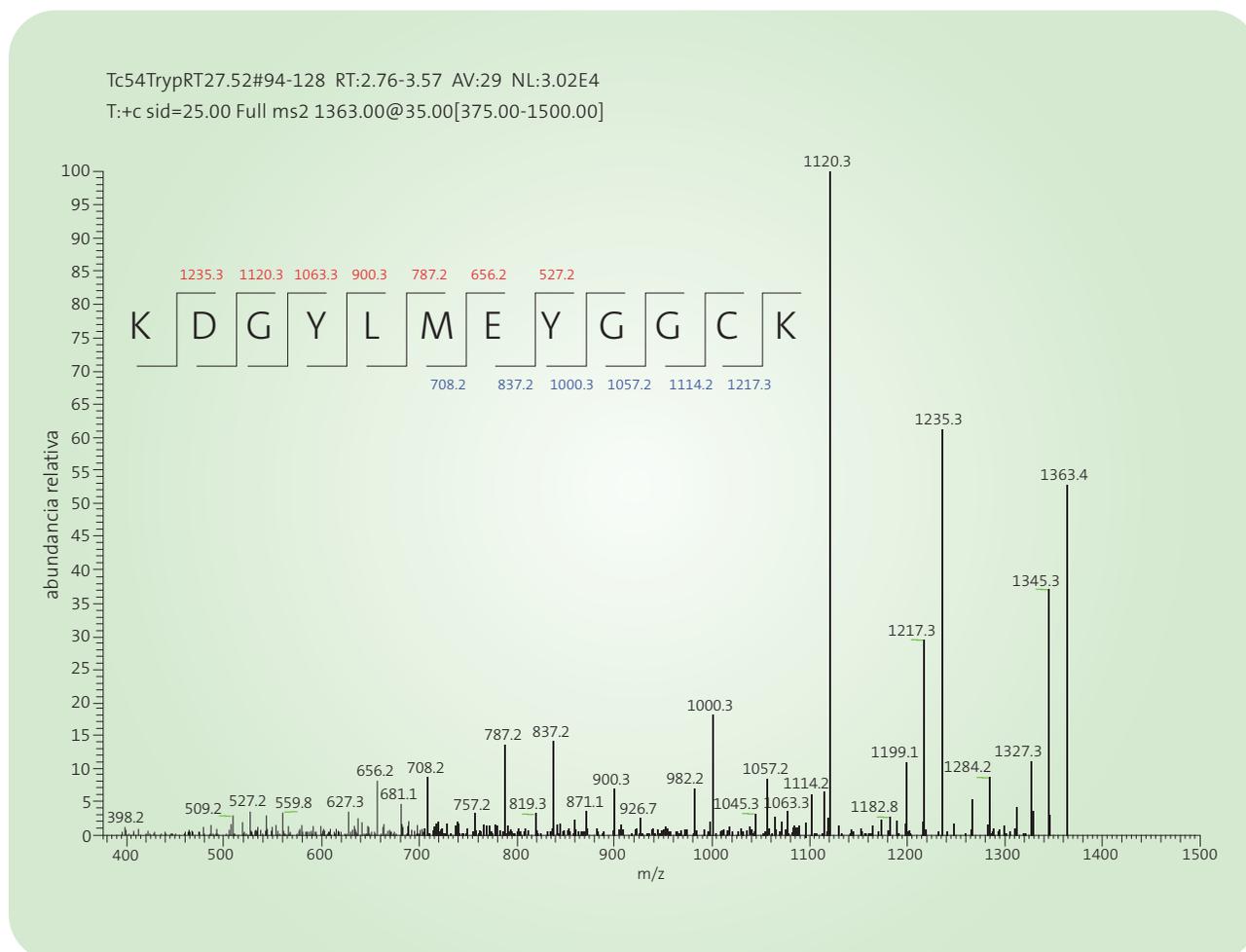
Finalmente, un elemento indispensable para el entendimiento de la función de las proteínas es el conocimiento de su estructura tridimensional. Sin embargo, pese al desarrollo de la bioinformática que permite la construcción de modelos estructurales *in silico* basado en homólogos estructurales, existe un gran número de proteínas con estructura tridimensionales no resueltas, como es el caso de las proteínas de membrana que representan un tercio del proteoma humano. Por esta razón, actualmente la comunidad científica internacional y las compañías de alta tecnología tienen como reto la integración de metodologías complejas y laboriosas como la difracción de rayos X y la resonancia magnética nuclear (RMN) al estudio de proteínas a gran escala. La posibilidad de integración de estas tecnologías no parece un proyecto a largo plazo, ya que los espectrómetros de masas híbridos, como los FT-ICR, utilizan súper magnetos para el análisis de proteínas con alta exactitud, acercándose de esta forma a la resolución de los equipos de RMN.

### La Unidad de Proteómica (UPRO) del Instituto de Biotecnología de la UNAM

La búsqueda de respuestas para preguntas biológicas complejas está directamente relacionada con el dominio, la aplicación y la disponibilidad de alta tecnología. En este sentido, la creación de unidades de servicio especializadas, a nivel regional o nacional, parece ser la mejor alternativa para la investigación científica y tecnológica en países en desarrollo. Las unidades de proteómica: a) son entidades técnica y operacionalmente elaboradas, b) por su natu-

raleza requieren grandes inversiones financieras, c) toman un tiempo razonable para su establecimiento y d) exigen recursos humanos altamente capacitados.

Dada la relevancia de las unidades de servicio en proteómica, tanto para la investigación básica como para la investigación biotecnológica, la comunidad de investigadores del Instituto de Biotecnología decidió la creación de la Unidad de Proteómica (UPRO-IBT) en junio de 2005. Esta unidad tiene como objetivo ofrecer servicios de análisis proteómico basados en la espectrometría de masas macromolecular. La UPRO es la primera y única unidad de servicios en su género a nivel nacional. Durante 2006 se analizaron más de ochocientos muestras de proteínas para instituciones de investigación pública y privada, así como para compañías particulares. Los servicios prestados por la UPRO incluyen determinación de masas moleculares de proteínas y péptidos, identificación de proteínas por MALDI-TOF y ESI-MS/MS, secuenciación parcial de proteínas desconocidas, determinación de modificaciones postraduccionales, determinación de sitios de corte específicos de nuevas enzimas e identificación de modificaciones sintéticas en proteínas. Adicionalmente, la UPRO es un centro proteómico de referencia a nivel nacional debido a la formación de recursos humanos (técnicos y de posgrado), al apoyo brindado a otros laboratorios de proteómica y a su empeño en promover y desarrollar la investigación proteómica en México (<http://www.smp.org.mx>). El significativo aumento de solicitudes de servicio es consecuencia del creciente interés de la comunidad científica nacional e internacional en dilucidar a nivel cualitativo y cuantitativo la expresión genética de los organismos, y especialmente la expresión diferencial de proteínas debida a distintas enfermedades, estrés u otras alteraciones físicoquímicas impuestas a los sistemas biológicos. La UPRO mantiene un esfuerzo constante por mejorar su plataforma analítica y, de esta forma, ofrecer servicios de primer nivel al creciente número de usuarios. ●



**Figura 2.**

Típico espectro de fragmentación por CID. El espectro se obtuvo en un espectrómetro de masas LC-MS (ionización ESI y analizador trampa de iones tridimensional). El ión precursor de 1363.4 Da ( $z=1^+$ ) se dividió en fragmentos de la serie *b* (azul) y de la serie *y* (rojo). Se muestra la secuencia de aminoácidos del ión precursor, la cual al ser sometida al programa BLAST, resulta en la secuencia de la toxina Tc54 del alacrán *Tityus cambridgei*.

### Bibliografía

- Batista, C. V. *et al.*, "Scorpion toxins from *Tityus cambridgei* that affect Na<sup>(+)</sup>-channels", en *Toxicon*, 40(5), 2002.
- Jacob, D. J. *et al.*, "The Complete genome and proteome of *Mycoplasma mobile*", en *Genome Res.*, 14(8), 2004.
- Olsen, J. V. *et al.*, "Global, in vivo, and site-specific phosphorylation dynamics in signaling networks", en *Cell*, 127(3), 2006.
- Siusdak, G., *Mass spectrometry in biotechnology*, MCC Press, San Diego, 2003.