Combinación de espectrometría de masas con punto isoeléctrico y determinación del aminoácido N-terminal de los péptidos para la mejor identificación de proteínas en estudios de proteómica

Autor Principal: Vladimir Besada[1], Aniel Sánchez[1], Yasset-Pérez Riverol[1]

Otros Autores: Lázaro Betancourt[1], Luis Javier González[1], Yassel Ramos[1], Jesús Noda[1], Félix Alvarez[1].

Colaboradores: Diogo Borges [2,3], Fabio C. S. Nogueira [5], Gilberto B. Domont[5], Felipe da Veiga Leprevost [3] , Felipe M. G. Franca [2] , Valmir C. Barbosa [2] , Paulo C. Carvalho [3], Alex Schmidt [7], Markus Müller [7], Jeovanis Gil [1], Roberto Vera [1], Gabriel Padron [1], Rui Wang [4], Juan Antonio Vizcaíno[4], Gabriel Duarte[5], David L. Tabb[6], Henning Hermjakob[4], Enrique Audain[8], Alelí Millan[1], Yoan J. Machado[8] .

Filiación:
[1]Centro de Ingeniería Genética y Biotecnología, Cuba;

[2]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, 21941-972 Rio de Janeiro, Brasil;

[3]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, 81350-010 Fiocruz, Paraná, Brasil;

[4]Proteomic Services, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK;

[5]Proteomics Unit, Institute of Chemistry, Federal University of Rio de Janeiro, 21941-909 Rio de Janeiro, Brasil;

[6]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA;

[7]Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU - 1, rue Michel Servet CH-1211 Geneva, Switzerland;

[8]Department of Proteomics, Center of Molecular Immunology, Cuba

Autor para la correspondencia:
Vladimir Besada Pérez: Centro de Ingeniería Genética y Biotecnología, ave 31 e/158 y 190, Playa, Ciudad Habana
vladimir.besada@cigb.edu.cu

1. Resumen:

La proteómica surge como consecuencia del desarrollo alcanzado por las técnicas suaves de ionización en la espectrometría de masas, fundamentalmente las conocidas como ESI y MALDI (del inglés electrospray ionization y matrix-assisted laser desorption ionization) que le permitió quedar directamente involucrada en la identificación de proteínas a partir de digestiones enzimáticas específicas. En estos veinte años de desarrollo la técnica evolucionó desde la identificación de una o varias proteínas aisladas en bandas de geles de electroforesis hasta el análisis masivo de mezclas complejas de péptidos y su identificación y cuantificación simultánea basada fundamentalmente en su espectro de masas MS/MS y la exactitud de la determinación de las masas moleculares tanto del ión precursor como de los iones fragmentos.

Sin embargo la elevada complejidad para cualquier especie en muestras de estudios de proteómica sobrepasa la capacidad de separación y el poder de detección de los métodos más avanzados de la cromatografía líquida multidimensional y la espectrometría de masas.

Por otra parte, los amplios rangos dinámicos de concentración de las distintas especies a analizar puede llegar a 10-12 órdenes lo que dificulta el análisis de las especies menos abundantes. La importancia de la función biológica de las proteínas no correlaciona con la abundancia por lo que en los estudios de proteómica tener acceso a proteínas cada vez menos abundantes es un reto omnipresente en el estado del arte.

De la cantidad enorme de péptidos que se generan de cualquier proteoma, al ionizarlos en el espectrómetro de masas solamente una pequeña porción de las señales se seleccionan para los experimentos de MS/MS y de estos una buena parte no generan espectros de masas con una calidad suficiente para realizar su identificación confiable. Esto es particularmente notorio para péptidos derivados de proteínas menos abundantes.

La aproximación de nuestro trabajo se basa en el uso combinado de la exactitud de la masa molecular, el punto isoeléctrico, la información del aminoácido ubicado en el extremo N, el tiempo de retención en la cromatografía de fase reversa, como criterios para la identificación de las proteínas, aún cuando no dispongamos de un espectro de masas MS/MS de elevada calidad. La combinación de estos datos con el uso de métodos de aislamiento selectivo de péptidos que permite analizar solamente péptidos con determinadas características en su secuencia también fue evaluada para incrementar la eficiencia de la identificación. Hasta el momento no se han usado estas características físico-químicas de los péptidos como criterio de identificación, el desarrollo fundamental ha sido desde el punto de vista tecnológico en la mejoría de la exactitud de los espectrómetros de masas, por lo que el uso combinado de estas características junto al espectro de masas, es totalmente novedoso.

Los resultados principales de este trabajo son:

1. La evaluación combinada del punto isoeléctrico, el aminoácido N-terminal, el tiempo de retención, la determinación de masas moleculares exactas por espectrometría de masas y los métodos de aislamiento selectivo, como criterio de identificación,

2. El desarrollo de un algoritmo y programa para un cálculo más exacto del punto isoeléctrico de los péptidos, basado en el diseño de máquinas de soporte vectorial, cálculo que se realiza masivamente para todos los péptidos identificados.

3. La determinación del aminoácido N-terminal de cada péptido modificado con isotiocianato de fenilo en su extremo N a partir de la fragmentación en fase gaseosa y la determinación del ión b1 en el espectro MS/MS, procedimiento que ocurre en condiciones de alto flujo.

4. El desarrollo de un programa para la identificación masiva de proteínas basados en la serie b1 del espectro de masas y la exactitud de la masa molecular, y se demostró un incremento de un 20% de nuevas identificaciones.

La **novedad** de este trabajo está en que:
1- se identificó la influencia que tiene el uso del punto isoeléctrico, el aminoácido N-terminal y el tiempo de retención en cromatografía de fase reversa, en combinación con la exactitud del espectro de masas, sobre la identificación de proteínas en estudios de proteómica,
2- se desarrolló un método y programa para la determinación masiva de punto isoeléctrico de los péptidos, que es más exacto que los métodos reportados,
3- se implementó un procedimiento para favorecer la determinación directa del aminoácido N-terminal de cada péptido a partir del fragmento b1 en el espectro de masas, posterior a su derivatización con PITC,
4- se desarrollaron varios programas de cómputo para la evaluación bioinformática de las bases de datos de péptidos y proteínas,
5- se elaboró un algoritmo y programa, que permite favorecer las identificaciones una vez se determine el aminoácido N-terminal, una secuencia parcial y su combinación con el espectro de masas,
6- se conformó una plataforma de trabajo que reduce los falsos positivos e incrementa el número de identificaciones en comparación con los programas más comúnmente usados.

El trabajo está avalado por cinco publicaciones que en el año de cada publicación tenían los siguientes factores de impacto: Analytical Chemistry (5.874; 5.695), Journal of Proteomics (4.878; 4.088), Bioinformatics (4.621). Todas son revistas de muy alto impacto dentro del campo de trabajo de proteómica.

2. Comunicación corta

La identificación de proteínas basada en la espectrometría de masas se basa en los espectros de masas del tipo MS/MS de los péptidos que la componen y la exactitud de la determinación de sus masas moleculares. Aún con los espectrómetros más modernos de alta velocidad de barrido, gran exactitud y sensibilidad, con frecuencia se requiere una reducción de la complejidad de la mezcla de péptidos que se estudian para facilitar su identificación en las bases

de datos de secuencias sin que se afecte la representatividad de las proteínas que le dieron origen.

El flujo de trabajo más empleado en los estudios de proteómica comprende la extracción de la mezcla de proteínas que se desea estudiar y su transformación en los péptidos que la componen mediante digestiones enzimáticas y su análisis de la mezcla resultante por cromatografía líquida acoplada a espectrometría de masas. La obtención de los espectros MS/MS es esencial para la identificación de proteínas en las bases de datos pues estos contienen información completa o parcial de la secuencia del péptido de interés.

Según el organismo de procedencia de los extractos proteicos este procesamiento de la muestra puede derivar en cientos de miles de péptidos que son, con los métodos cromatográficos y electroforéticos actuales, imposibles de resolver y por tanto decenas de ellos llegan simultáneamente al detector del espectrómetro de masas. De la cantidad enorme de péptidos que se generan de cualquier proteoma, solo una parte se seleccionan para los experimentos MS/MS y de estos una buena parte no generan espectros de masas con una elevada calidad que garanticen una identificación confiable, por lo que al final menos de un 50% de los péptidos presentes se pueden identificar.

En este trabajo exponemos diversos procedimientos que evaluamos e introdujimos en nuestro grupo para mejorar la identificación de proteínas a partir no solo de aumentar el número de péptidos identificados, sino también de la reducción de los falsos positivos. Adicional al espectro MS/MS y al uso de la exactitud en la determinación de las masas moleculares de los péptidos se propuso incluir como criterios identificativos otras propiedades físico-químicas que usualmente se obtienen durante el procesamiento previo de la muestra y se desechan o no se consideran de manera rutinaria por los principales motores de búsqueda utilizados en la actualidad para la identificación de proteínas en las bases de datos de secuencia. Estas propiedades son el punto isoeléctrico, el tiempo de retención como criterios para la identificación de las proteínas. Por otra parte, en esta estrategia de trabajo se propone la modificación química de los péptidos con fenilisotiocianato de fenilo (PITC) para favorecer la formación en los espectros MS/MS de la serie b1 que suministra información del amino ácido N-terminal y así aun cuando no dispongamos de un espectro de masas MS/MS de alta calidad, se pueden realizar identificaciones confiables de proteínas.

Por otra parte, el incluir varias propiedades físico-químicas en combinación con el asilamiento selectivo de péptidos como criterio identificación permite incrementar el número de péptidos únicos que son los que realmente contribuyen de manera inequívoca a la identificación confiable de las proteínas y a la selectividad del método implementado.

El aislamiento selectivo de péptidos desarrollados en nuestro grupo permite aislar péptidos a partir de mezclas complejas que tienen rasgos comunes en su secuencia. Este proceder permite filtrar las bases de datos de péptidos al incluir los que tengan las características deseadas. De esta manera se reduce el espacio de búsqueda y se incrementa el número de péptidos únicos sin que se afecte a su vez la representatividad de las proteínas que le dieron origen.

Cuando se combinan estos criterios con los métodos de aislamiento selectivo, el número de péptidos únicos y las proteínas identificadas se incrementan sustancialmente. Los resultados demostraron que es posible la identificación con alta confiabilidad del 93% de las proteínas identificadas.

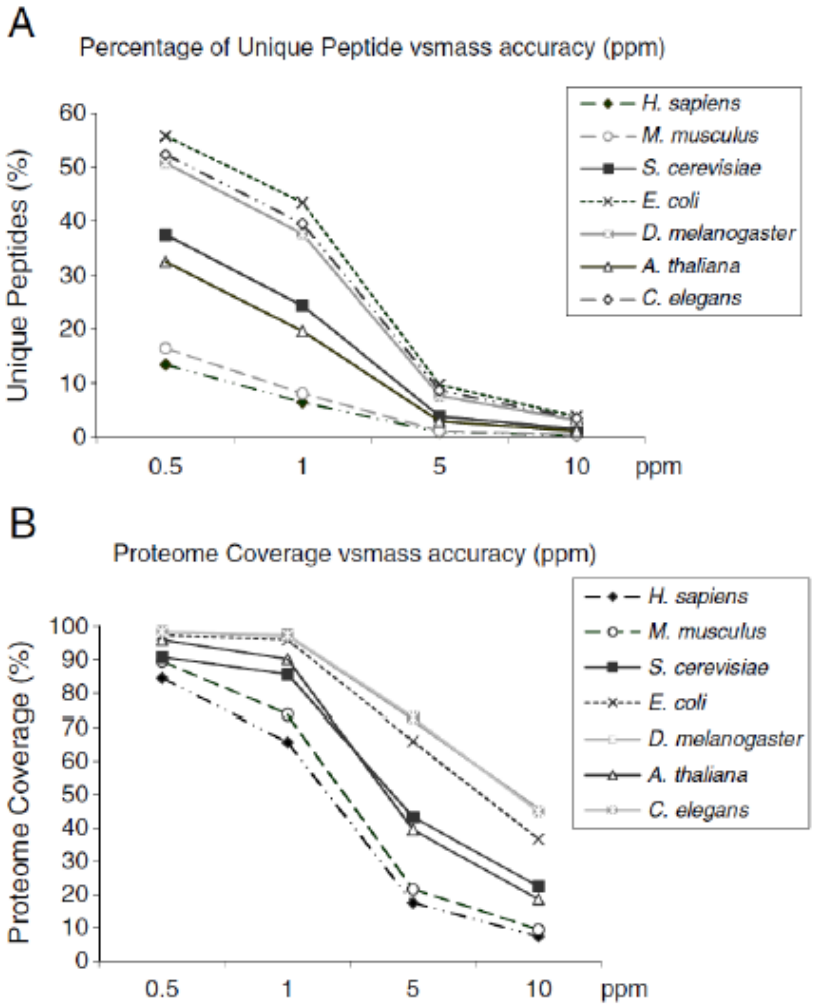
Para lograr estas capacidades trabajamos en varios sentidos:
1- Evaluar *in silico* la influencia que pueden tener en la identificación de proteínas en las bases de datos varias propiedades físico-químicas, fundamentalmente el punto isoeléctrico, el amino ácido N-terminal y el tiempo de retención.

2- Desarrollar un software que posibilite una determinación teórica confiable del punto isoeléctrico basado en máquinas de soporte vectorial,

3- Establecer un método de derivatización del N-terminal de péptidos con PITC y estudiar la fragmentación de los péptidos modificados, con el objetivo de determinar en los espectros MS/MS el amino ácido N-terminal de los péptidos,

4- Desarrollar un programa para la identificación de péptidos empleando la exactitud de masas y el amino ácido N-terminal.

El uso de la masa molecular únicamente para lograr una identificación de péptidos en las bases de datos de proteínas no es confiable pues el número de péptidos únicos siguiendo nada más este criterio, es despreciable. En los estudios de proteómica el tiempo de retención de los péptidos durante la corrida cromatográfica en la columna de fase reversa es uno de las propiedades más usadas como criterio de eliminación de identificaciones falsas-positivas, es de hecho uno de los elementos que más incrementa el porciento de péptidos únicos.
De modo general aunque se nota un aumento del número de péptidos únicos con la inclusión del tiempo de retención, no fue un cambió tan relevante, aun usando el aislamiento selectivo de péptidos, como la inclusión del punto isoeléctrico y el N-terminal. El uso de los cuatro parámetros (masa, p*I*, tiempo de retención y N-terminal) combinado con algún método de captura selectiva permite alcanzar más de un 90% de péptidos únicos. Los resultados muestran que el incremento de péptidos únicos es más apreciable para proteomas complejos y el uso de los métodos de aislamiento contribuyen en casi el doble en relación con el análisis de la mezcla completa de péptidos.
Los resultados mostraron que la cobertura del proteoma puede ser tan alta como el 60% u 80% para todos los organismos analizados a exactitudes de 0.5 y 1 ppm, respectivamente. Entiéndase por cobertura del proteoma el % de proteínas que pueden ser identificadas de las que están almacenadas en las bases de datos de secuencias.



(A) Porciento de péptidos únicos del proteoma completo de diferentes organismos digeridos *in silico* con tripsina considerando solo su masa molecular. (B) Cobertura teórica del proteoma de diferentes organismos si solo se consideran péptidos únicos. Los valores del eje X corresponden con el rango de ppm evaluado (0.5, 1, 5 and 10).

Esta evaluación *in silico* permitió dirigir nuestra investigación hacia el posible uso de estos parámetros en la identificación de péptidos teniendo en cuenta que algunos de ellos es posible predecirlo o determinarlo empíricamente durante el fraccionamiento propio de las mezclas de péptidos. Fue posible predecir que considerar el punto isoeléctrico (p*I*) y el aminoácido N-terminal debía dar los mejores resultados. Esta primera etapa de la investigación está publicada en "*In silico* analysis of accurate proteomics, complemented by selective isolation of peptides. *J Proteomics. 2011; 74(10):2071-82".*

El segundo resultado es la identificación de descriptores de los aminoácidos que permiten relacionar sus propiedades con el punto isoeléctrico. A partir de estos descriptores se desarrolló un programa que permite estimar con mayor exactitud el punto isoeléctrico de los péptidos, incluso de numerosos péptidos simultáneamente. Esta determinación teórica del punto isoeléctrico y la comparación con valores experimentales posibilita usar este parámetro como criterio para descartar o incluir los péptidos identificados.

Varios descriptores moleculares de los aminoácidos fueron considerados como son el índice de refracción, la polarizabilidad, el área superficial, entre otras muchas propiedades físico-químicas y biológicas. Para cada péptido se computaron los 543 descriptores existentes en la base de datos AAindex, con la siguiente expresión matemática PD = (ΣAD)/NA donde PD es el promedio de los descriptores de todos los aminoácidos de AAindex. También incluimos los valores de p*I* predichos por otros autores.

El nuevo modelo desarrollado basado en SVM (del inglés "support vector machine") fue entrenado y probado con 7391 péptidos derivados de un estudio de proteómica realizado en células Kc167 de *Drosophila melanogaster*. Los péptidos fueron previamente fraccionados por electroforesis OFFGEL y luego cada fracción fue analizada en un espectrómetro de masas LTQ-FT-ICR equipado con ESI. Se arribó a un modelo que permitió mejorar sustancialmente el cálculo de punto isoeléctrico comparado a otros dos modelos existentes, con un error no mayor de 0.32 unidades de p*I* y una correlación de 0.98 entre los valores calculados y los determinados experimentalmente. El programa permite el cálculo de gran cantidad de péptidos simultáneamente, lo que es una herramienta importante en los estudios de proteómica, pues hemos demostrado que el punto isoeléctrico puede ser un criterio de relevancia para descartar identificaciones falsas positivas. Este trabajo está publicado en "Isoelectric point optimization using peptide descriptors and support vector machines. *J Proteomics. 2012 Apr 3;75(7):2269-74".*

Para lograr determinar el residuo presente en el extremo N de cada péptido profundizamos en un procedimiento previamente descrito por Gaskell y cols. [i] en el cual los péptidos derivatizados con isotiocianato de fenilo (PITC) eran capaces de generar en fase gaseosa (dentro del espectrómetro de masas) y con poca energía, iones del tipo b1 que aportan directamente la información del aminoácido del extremo N. Investigamos entonces las mejores condiciones para la obtención de estos derivados y las energías de colisión para iones de diferente carga, así como la influencia de varios aminoácidos en la formación de este ión.

En nuestro trabajo se evaluaron también los péptidos derivados de las proteínas Albúmina de suero bovino (BSA) y Estreptoquinasa (STK). Se evaluaron los péptidos luego de usar el esquema de fraccionamiento por intercambiador catiónico fuerte (aislando péptidos con cantidad de argininas e histidinas mayor que 1), y se pudo demostrar la conveniencia de este procedimiento para estudios de proteómica que incluya fraccionamiento por cargas. Otros métodos de aislamiento o fraccionamiento también fueron evaluados.

Como prueba de concepto el procedimiento se aplicó al análisis de proteínas citosólicas de *E. coli*. Se realizaron dos experimentos para el análisis de los péptidos derivatizados con PITC, a) sin aislamiento selectivo y b) despues del aislamiento selectivo de péptidos multicargados. Un aspecto interesante es el incremento observado en la cantidad de péptidos únicos identificados cuando se incluye la información del aminoácido N-terminal además del espectro MS/MS. Este comportamiento se observó in silico para los proteomas teóricos de *E.coli* y *Homo sapiens*.

En proteomas poco complejos como el de E.coli que contiene unas 4300 proteínas es posible identificar cerca del 80 de las proteínas a partir de los péptidos únicos, cuando se trabaja con espectrómetros de masas con menos de 5 ppm de exactitud, con péptidos trípticos derivatizados con PITC y aislando los péptidos del tipo RH. Este trabajo se publicó en "Evaluation of Phenylthiocarbamoyl-Derivatized Peptides by Electrospray Ionization Mass Spectrometry: Selective Isolation and Analysis of Modified Multiply Charged Peptides for Liquid Chromatography-Tandem Mass Spectrometry Experiments. *Anal Chem. 2010, 82, 8492–8501".*

El reto en este momento fue establecer un procedimiento completo, incluyendo una herramienta bioinformática, que permitiera la identificación en bases de datos de secuencias de proteínas a partir de los ficheros completos de espectros obtenidos en corridas cromatográficas de LC-MS/MS.

Al surgir esta posibilidad nos interesamos en generar herramientas bioinformáticas que permitieran la identificación de proteínas en bases de datos, que le diera más puntuación o relevancia a la identificación del residuo presente en el extremo N. Los programas hasta el momento existentes no le dan tanta relevancia a esto porque normalmente no se tiene esa información, la podemos extraer ahora a partir de los péptidos derivatizados con PITC.

El programa incluye un procedimiento para reconocer la serie b1 a partir de un ordenamiento que realiza por intensidades de los iones fragmentos de bajas masas. En esta estrategia de trabajo permitimos la generación de otros iones fragmentos adicionales a la serie b1 y el programa es capaz de reconocer segmentos de secuencias de aminoácidos en la región media del espectro, de modo que tenemos una variante intermedia entre la búsqueda en bases de datos por comparación con espectros teóricos y la secuenciación *de novo*.

Al determinar cuánto mejora la inclusión del residuo N-terminal únicamente en dos proteomas (*E.coli y H.sapiens*) se observa un incremento significativo de los péptidos únicos y de la cobertura de secuencia de las proteínas. Estos resultados aparecen reflejados en dos publicaciones, la primera en "Effectively addressing complex proteomic search spaces with peptide spectrum matching. *Bioinformatics. 2013 Vol. 29 no. 10, 1343–1344"* que trabaja con espacios de busqueda restringidos y la segunda "HI-Bone: A Scoring System for Identifying Phenylisothiocyanate-Derivatized Peptides Based on Precursor Mass and High Intensity Fragment Ions. *Anal Chem. 2013 Apr 2;85(7):3515-20"* que es el desarrollo del programa HI-bone empleando el aminoácido del N-terminal y los métodos de captura selectiva.

En *E.coli* se demostró un incremento del 20% del número de identificaciones de proteínas con relación al uso de los dos software más comúnmente usados.

2.1 Aporte científico y metodológico:

El trabajo se basa en la potencialidad que representa el uso combinado de varias propiedades físico-químicas como son el punto isoeléctrico, la exactitud de la masa molecular, el aminoácido N-terminal de los péptidos, y el tiempo de retención en la identificación de las proteínas en las bases de datos para estudios de alto flujo. Esto no había sido previamente descrito, solo el tiempo

de retención ha sido empleado para estudios de validación (no para identificación de proteínas) y hay programas para su predicción.
La información de estas propiedades puede obtenerse de manera rutinaria durante las etapas de procesamiento de la muestra previo al análisis por espectrometría de masas y por lo general estas propiedades no son consideradas como criterios de identificación de los péptidos en bases de datos de secuencias utilizando la espectrometría de masas como herramienta analítica.
Por otra parte, en esta metodología se demuestra que no es necesario obtener espectros MS/MS de elevada calidad para obtener una identificación confiable pues solo es necesario obtener información del aminoácido que está presente en el extremo N-terminal al detectar la serie b1 que es muy favorecida para los péptidos modificados con fenilisotiocianato. Este apecto es importante para la identificación de péptidos poco abundantes o que por las características particulares de sus secuencia produzcan espectros MS/MS de poca calidad.
Los procedimientos desarrollados forman parte de la plataforma de trabajo del CIGB para estudios de proteómica. Los resultados que se presentan en este trabajo corresponden a desarrollos metodológicos que tuvieron lugar fundamentalmente entre 2009 y 2013. Los resultados aquí mostrados no son parte de anteriores Premios de la Academia de Ciencias de Cuba.

Aporte económico: No hay un aporte económico directo. Se trata de un desarrollo metodológico que posibilita una identificación más confiable de péptidos y proteínas, que ha sido probado en sistemas proteicos simples, como digestiones peptídicas de proteínas puras y extractos proteicos de la bacteria *E.coli*. Se evita la identificación basada únicamente en la secuencia de los péptidos extraída de los espectros MS/MS que comúnmente está limitada por el hecho de que este proceso no ocurre siempre con eficiencia dentro del espectrómetro cuando se trabaja en alto flujo.

Importancia práctica del trabajo:
1- Se caracterizaron las bases de datos de diversos organismos y se determinaron dos de las propiedades de mayor influencia en el proceso de identificación de proteínas: punto isoeléctrico y N-terminal,
2- la aplicación de este método demostró in silico un incremento de entre el 60 y más del 90% de péptidos único según el proteoma en estudio,
3- Se aplicó el cálculo de punto isoeléctrico masivo a un proyecto de Drosophila melanogaster, que resultó en una determinación más exacta con respecto a lo teórico,
4- Se demostró el incremento en proteínas identificadas en un extracto proteico de *E.coli*.
5- Se cuenta con una nueva metodología de trabajo que permitirá mejorar las identificaciones de proteínas a partir de un incremento en el número de espectros de masas coincidentes con espectros teóricos y de una disminución de identificaciones falsas positivas.

Referencias

[1] Summerfield, S. G.; Bolgar, M. S.; Gaskell, S. J. J. Mass Spectrom. 1997, 32, 225–231.
[2] Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, et al. Electrophoresis 1993;14:1023–31.
[3] Cargile BJ, Sevinsky JR, Essader AS, Eu JP, Stephenson Jr JL. Electrophoresis 2008;29:2768–78.
[4] Pérez-Riverol Y, Sánchez A, Ramos Y, Schmidt A, Müller M, Betancourt L, González LJ, Vera R, Padron G, Besada V. J Proteomics. 2011 Sep 6;74(10):2071-82.
[5] Zimmerman JM, Eliezer N, Simha R. J. Theor Biol 1968;21:170–201.
[6] Liu HX, Zhang RS, Yao XJ, Liu MC, Hu ZD, Fan BT. J Chem Inf Comput Sci 2004;44:161–7.
Referencias: