Sesión de bioinformática

Coordinadores: [1]Salvador Martínez de Bartolomé, [2]Pedro Navarro

[1] ProteoRed, Centro Nacional de Biotecnología - CSIC, Cantoblanco, Madrid, Spain
[2] Instituto de Biología de Sistemas Moleculares (IMSB), Instituto Federal Suizo de Tecnología (ETH), Zürich, Suiza

Resumen

En las pasadas II Jornadas de Jóvenes Investigadores en Proteómica, celebradas en Córdoba, tuvo lugar una fructífera sesión de Bioinformática, organizada como una mesa redonda, en la que diversos temas relacionados con el análisis bioinformático en Proteómica, votados por los socios de la SEProt, fueron introducidos por expertos de forma que se produjera un debate alrededor de ellos.

La sesión de bioinformática de las II Jornadas de Jóvenes Investigadores en Proteómica fue organizada en un formato de mesa redonda que permitiera participar a todos los asistentes de una forma activa. Para que la sesión transcurriera de la manera más provechosa posible, previamente a las jornadas se organizó una votación online entre los miembros de la SEProt de los posibles temas de mayor interés, como son: herramientas de biología de sistemas, análisis de resultados de proteómica cuantitativa, software libre en proteómica, combinación de motores de búsqueda, estándares de representación de experimentos y estándares para almacenamiento e intercambio de datos proteómicos. Teniendo en cuenta el interés mostrado en las votaciones por cada uno de estos temas, se trató invitar a diferentes jóvenes investigadores de la SEProt, expertos conocedores de alguno o algunos de dichos temas y así poder responder a cualquier pregunta que pudiera surgir en torno a ellos.

Alberto Medina nos presentó una introducción a las FPA (Functional Proteomics Annotations), anotaciones que proveen a los investigadores de algunas claves biológicas relacionadas con la proteína o péptido que contenga este tipo de anotación. Las FPA pueden ser producidas de manera manual, o de manera automática: ambos métodos producen desafíos interesantes, como planteó Alberto en su presentación del tema, ya que existe un gran número de bases de datos que almacenan información de Proteómica, con multitud de referencias cruzadas entre ellas. Además, cada base de datos utiliza índices propios, con lo que el intercambio entre bases de datos exige un mapeo entre palabras clave e índices que no siempre es trivial. Alberto se pregunta si es realmente posible obtener los mismos datos desde sitios diferentes, y si existen flujos de trabajo que permitan esta recuperación de la información. Uno de los sistemas que permite una recuperación automática de FPAs es PIKE [1], un sistema web que permite una comprobación de la información recuperada, así como un análisis de datos mediante *clustering*, y análisis mediante las FPA de dominios, pathways e interacciones entre proteínas, pero Alberto plantea a la audiencia si realmente podemos confiar en este tipo de sistemas, y cómo podemos obtener de ellos datos fiables (o medir su fiabilidad), ya que existe una gran dependencia de las fuentes de bases de datos desde donde éstos son adquiridos.

Juan Antonio Vizcaíno fue invitado como experto en estándares en Proteómica. Actualmente, la enorme cantidad de información accesible desde PubMed no está siendo aprovechada al máximo debido a la dificultad de compartir información en diferentes formatos, lo que dificulta las comparaciones inter-experimentales e inter-laboratorios. El uso de estándares puede permitirnos comparaciones de metodología, de protocolos y de instrumentos, aunque la situación actual no es la mejor posible, ya que aún existe una cierta desconexión entre los diversos fabricantes de equipos de espectrometría de masas (Agilent, Waters, AB, Thermo…), los motores de búsqueda más

utilizados (Phenyx, Mascot, ProteinPilot, SEQUEST…), y los repositorios de información tales como PeptideAtlas [2], theGPMdb [3] y PRIDE [4], de forma que aún se utilizan diferentes formatos, tanto en la entrada de datos, como en la obtención de resultados. Afortunadamente, esta situación está cambiando poco a poco, de forma que cada uno de los organismos involucrados es más consciente de la necesidad de utilizar estándares, siguiendo tres principios fundamentales: la información debe estar disponible, con lo que deben existir métodos y/o herramientas que permitan acceder a toda la información que corresponda a, por ejemplo, una proteína o un experimento concreto (la base de datos PRIDE es un buen ejemplo de almacenamiento y disponibilidad de datos proteómicos basado en un estándar), la información debe poder ser recuperada fácilmente, y para ello deben utilizarse formatos estándar que garanticen una comprensibilidad universal de los datos y, además, la información generada y almacenada debe ser suficientemente explicativa, para lo cual se define la mínima información necesaria para describir un experimento proteómico (MIAPE [5]). La iniciativa HUPO-PSI [6] trata de ayudar en el cumplimiento de estas dos últimas premisas, desarrollando formatos estándar, representaciones de datos y anotaciones estándar, definiendo las directrices MIAPE e involucrando en estos desarrollos a productores de datos, proveedores de bases de datos, desarrolladores de software y editores de las principales revistas científicas y proteómicas. Hasta el momento, esta iniciativa ha desarrollado de manera exitosa documentos de especificación MIAPEs y conjuntos de vocabularios controlados (ontologías) en Proteómica, así como diversos estándares de datos, como el mzML [7], que reúne la misma información sobre la adquisición de datos en espectrometría de masas contenida en los formatos, más antiguos, mzData y mzXML [8], añadiendo además un vocabulario controlado. Otro formato recientemente desarrollado, el mzIdentML, está destinado a la expresión de resultados en motores de búsqueda, lo que permite la comparación de resultados de los motores de búsqueda más conocidos, aunque actualmente la información contenida en este formato sea exclusivamente cualitativa. En el momento se está trabajando duramente en encontrar una buena solución para expresar resultados de Proteómica cuantitativa en un formato estándar (mzQuantML).

El ponente Alex Campos, experto en software libre aplicado en Proteómica, resumió un número interesante de aplicaciones gratuitas de conversión de formatos propietarios a formatos estándar, tales como ProteoWizard [9]. Además, propuso algunas alternativas de software que permiten realizar extracción de picos, y cuantificación, tanto basada en marcaje con isótopos estables como sin marcar (OpenMS [10], SuperHirn [11]), y preguntó a la audiencia por los diversos programas que utilizaban habitualmente, generando un interesante debate sobre las diferentes alternativas entre los laboratorios de la SEProt, problemas de conversión de datos, y métodos para intercambiar información de resultados de Proteómica cualitativa y cuantitativa entre laboratorios colaboradores.

Por último, Marco Trevisan nos recordó que los formatos XML no deben ser utilizados en general como métodos de almacenamiento de datos, práctica habitual en muchos laboratorios, y que el almacenamiento de datos debe ser realizado en formatos específicos de almacenamiento de datos, como son las bases de datos. En la discusión posterior se comentó que el formato XML nació de facto como un formato de intercambio de información, y nunca como un método de almacenamiento, con lo que debería respetarse esta filosofía de trabajo.

Los organizadores de esta sesión y los ponentes de la misma queremos agradecer a todos los asistentes a la sesión su alta participación durante la mesa redonda, y deseamos que haya sido provechosa para sus intereses. Deseamos así mismo que el nivel científico mostrado durante todas las Jornadas se mantenga en próximas citas, y felicitamos a los organizadores de Córdoba por su excelente trabajo. ¡Hasta la próxima!

Referencias

[1] Medina-Aunon JA, Paradela A, Macht M, Thiele H, Corthals G y Albar JP. PIKE: discovering biological information from proteomics data. PROTEOMICS. 2010 Sep;10(18):3262-71.

[2] Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, et al. The PeptideAtlas project. Nucl. Acids Res. 2006;34:D655-58.

[3] Craig R, Cortens JC, Fenyo D y Beavis RC. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 2006;5:1843-9.

[4] Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucl. Acids Res. 2006;34:D659-63.

[5] Taylor C, Paton N, Lilley K, Binz P-A, Julian R, Jones A, et al. The minimum information about a proteomics experiment (MIAPE). Nature Biotechnology 2007;25:887-93.

[6] Orchard S, Hermjakob H y Apweiler R. The proteomics standards initiative. PROTEOMICS 2003;3:1374-6.

[7] Deutsch E. mzML: A single, unifying data format for mass spectrometer output. PROTEOMICS 2008;8:2776-77.

[8] Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 2004;22:1459-66.

[9] Kessner D, Chambers M, Burke R, Agus D y Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008;24:2534-6.

[10] Kohlbacher O, Reinert K, Gropl C, Lange E, Pfeifer N, Schulz-Trieglaff O, et al. TOPP--the OpenMS proteomics pipeline. Bioinformatics 2007;23:e191-7.

[11] Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, et al. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. PROTEOMICS 2007;7:3470-80.