

Título:

Modelo estadístico de aplicabilidad general para el análisis de datos de proteómica cuantitativa de alto rendimiento

Resumen:

Investigadores del CNIC han desarrollado y validado un modelo estadístico general, denominado modelo WSPP, para el análisis estadístico de resultados de proteómica cuantitativa. El modelo describe de forma muy precisa la variabilidad técnica de los datos de los métodos más comunes de proteómica cuantitativa mediante marcaje con isótopos estables (SIL). Además, el modelo permite la comparación sistemática y la integración de datos obtenidos a partir de diferentes experimentos.

El modelo WSPP considera por separado la variabilidad introducida durante (i) la extracción y manipulación de las proteínas, (ii) la generación de péptidos a partir de sus correspondientes proteínas y su marcaje, y (iii) la generación de información cuantitativa mediante espectrometría de masas (EM). Cada uno de estos pasos se analiza de modo independiente, proporcionando así por primera vez la flexibilidad necesaria para tener en cuenta la distinta naturaleza de las fuentes de error asociadas a los diferentes métodos. La validez general del modelo se demostró mediante la confrontación de 48 distribuciones experimentales frente a 18 hipótesis nulas diferentes.

El modelo WSPP constituye el primer estándar integrado de validez general para el análisis de datos de proteómica cuantitativa obtenidos mediante marcaje con isótopos estables.

Aspectos innovadores:

La proteómica basada en espectrometría de masas (EM) permite la identificación y cuantificación relativa de miles de proteínas en un único ensayo. Sin embargo, los modelos existentes son altamente específicos para cada método SIL y cada espectrómetro de masas, por lo que son inadecuados para el examen de datos procedentes de diferentes laboratorios, la determinación de la calidad experimental en base a criterios unificados, la comparación e integración de múltiples mediciones, o la interpretación como un todo del conjunto completo de los resultados experimentales obtenidos a partir de diferentes estrategias SIL. Por otra parte, la mayoría de los modelos y las pruebas de significación estadística existentes se basan en supuestos de normalidad que no han sido demostrados a pesar del hecho de que la heterogeneidad de la varianza ha sido documentada en todos los métodos SIL. Estas técnicas generan información cuantitativa a nivel de péptido y la ausencia de modelos generales hace muy subjetiva la elección del método adecuado para integrar esta información y obtener resultados a nivel de proteína. Este problema se agrava aún más por el submuestreo que caracteriza los análisis de EM basados en SIL: el número de péptidos que cuantifican una proteína es variable y no se puede controlar entre diferentes experimentos; este hecho no es nada trivial y ha complicado en gran medida el modelado matemático de este tipo de datos.

El modelo WSPP constituye el primer marco estadístico de aplicación general para el análisis de datos generados con tecnologías SIL-EM. El modelo proporciona una descripción detallada del comportamiento de la varianza a todos los niveles, capturando por separado las diferentes fuentes de error específicas de cada método SIL y técnica de EM, y demostrando que en todos los casos las distribuciones de error se modelan con precisión en todos los niveles.

No se ha formulado previamente ningún modelo para datos de proteómica cuantitativa que descomponga la varianza técnica en dos o más componentes. Además, en el modelo WSPP la integración de datos se lleva a cabo siguiendo la teoría de propagación de errores, de modo que el peso estadístico con el que cada valor contribuye a las medidas finales es exactamente el inverso de su varianza local. Por otro lado, el modelo WSPP es estadísticamente robusto porque asigna con exactitud una varianza local a todos los valores cuantitativos, basándose solo en cuatro parámetros que se estiman a partir del análisis conjunto de toda la colección de datos.

Este modelo estadístico también proporciona por primera vez un marco para comparar e integrar los resultados obtenidos mediante diferentes aproximaciones cuantitativas en diferentes tipos de espectrómetros de masas.

Ventajas competitivas:

El modelo permite una integración coherente de la información cuantitativa, teniendo en cuenta todas las fuentes de error y de acuerdo a la teoría de propagación de errores, de manera que la varianza específica de cada valor en cualquier nivel se estima con precisión. El modelo resuelve eficazmente el problema de submuestreo, proporcionando un marco adecuado para analizar los datos en cada nivel usando distribuciones normales únicas. Además, permite la comparación e integración de los resultados obtenidos mediante diferentes técnicas SIL, de manera que el control total sobre la varianza se mantiene en los datos integrados, abriendo así la posibilidad de realizar más integraciones en niveles superiores.

Este marco estadístico resuelve eficazmente los problemas de heterogeneidad de varianza, la integración de datos y el submuestreo y proporciona un método estadísticamente sólido para analizar la calidad de experimentos cuantitativos y detectar desviaciones experimentales y para integrar resultados obtenidos mediante diferentes técnicas cuantitativas. Este algoritmo estadístico ofrece por primera vez un marco estadístico válido para analizar cualquier tipo de datos de proteómica cuantitativa.

Palabras clave: proteómica cuantitativa, marcaje con isótopos estables, análisis estadístico

Contacto:
OTRI-CNIC
proyectos_otri@cnic.es