Modelización estadística del proteoma completo con células humanas en tejido normal y tumoral

Raquel Gavidia[1], Carmen Armero[1], M. Luz Valero[2], Manuel Sánchez[2] y Laia Bernet[3]

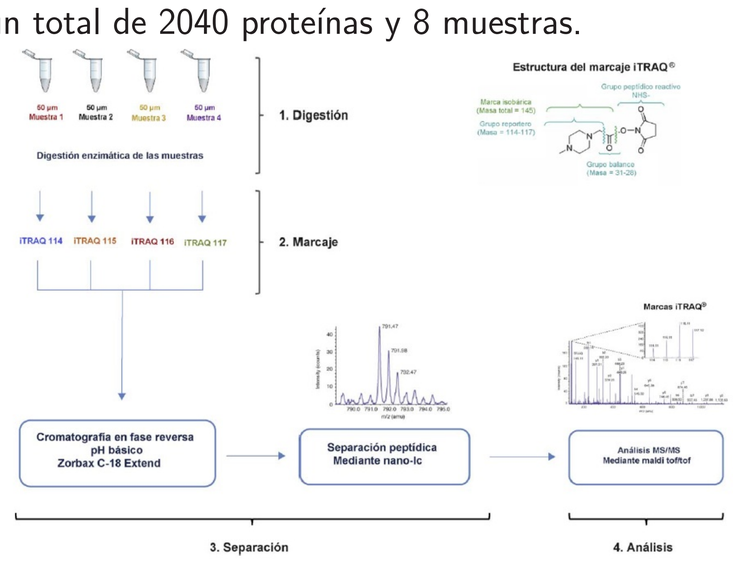
[1]Departament d'Estadística i Investigació Operativa, Universitat de València
[2]Secció de Proteòmica del SCSIE, Universitat de València
[3]Hospital Lluis Alcanyís, Xàtiva

1. Introducción

- Las investigaciones proteómicas tratan de identificar y cuantificar las variaciones de abundancia observadas en una proteína o grupos de proteínas en individuos sanos y enfermos con el fin de descubrir biomarcadores de pronóstico o diagnóstico.
- Cualquier enfermedad conlleva la alteración de determinadas proteínas, que pueden provocar a su vez modificaciones sobre otras proteínas produciendo un efecto cascada de alteraciones en las que pueden verse implicadas un elevado número de proteínas.
- Objetivo del estudio:
  - Identificar aquellas proteínas que diferencian el tejido sano y enfermo en mujeres con cáncer de mama.

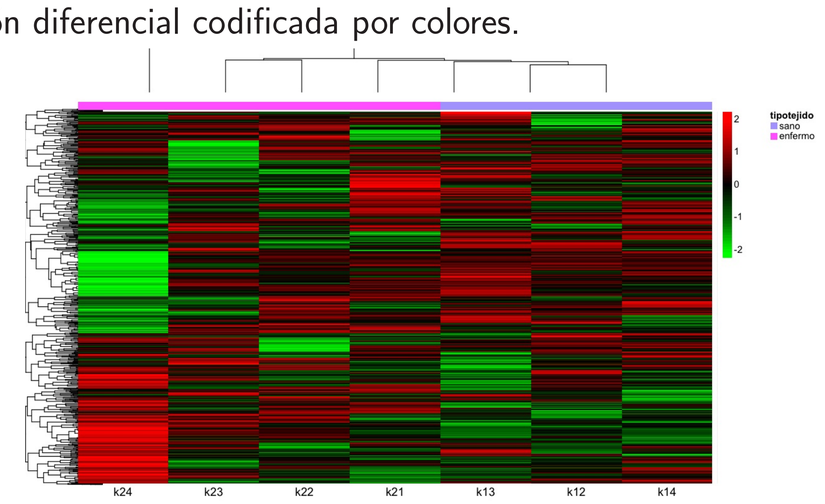
2. El problema

- Descripción del banco de datos:
  - Estudio realizado en cuatro mujeres enfermas de cáncer de mama.
  - A una misma paciente se le extrae una muestra de tejido sano y enfermo.
  - Asumimos independencia debido al pequeño número de tejidos disponibles.
- La técnica proteómica utilizada es el marcaje i-TRAQ.
  - Cuantifica y observa modificaciones proteicas mediante un marcaje químico.
  - Marca como máximo 8 muestras por experimento.
  - Tenemos un total de 2040 proteínas y 8 muestras.



3. La base de datos

- Presentación de la base de datos en términos matriciales:
  - Las filas corresponden a las proteínas y las columnas a las muestras (reconvertidas en 7).
  - Datos normalizados de la expresión diferencial de cada proteína correspondientes al tejido sano y enfermo de las cuatro mujeres para evitar problemas de escala .
  - Expresión diferencial codificada por colores.

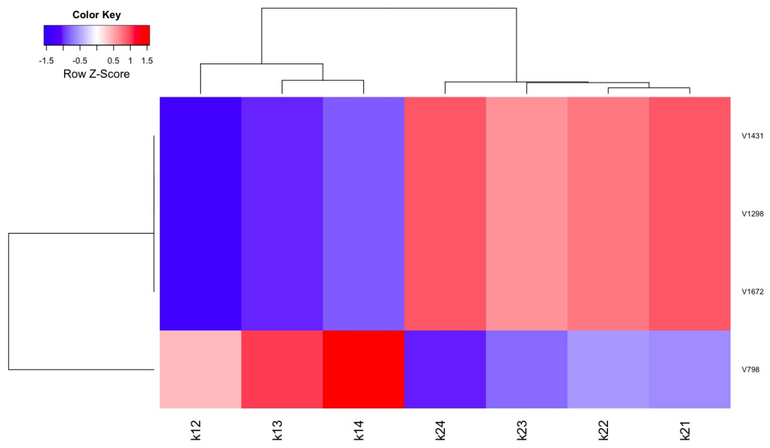


4. Líneas de trabajo

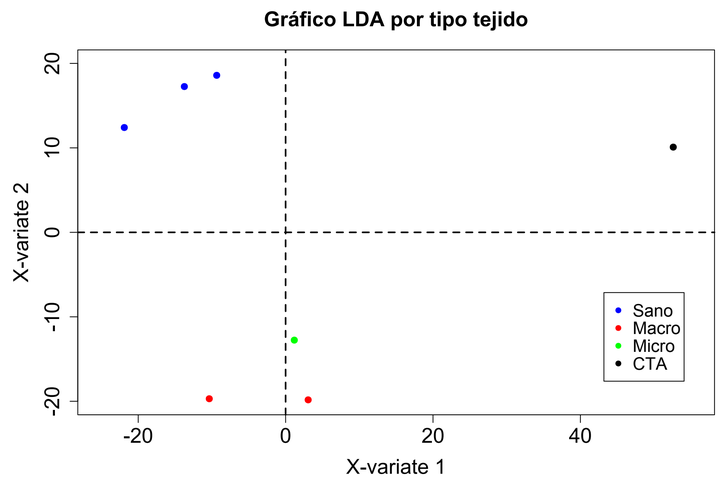
- 4.1 Comparaciones múltiples de la expresión diferencial media de las proteínas en tejidos sanos y enfermos a través del False Discovery Rate (FDR).
- 4.2 Modelos de regresión logística con penalización Lasso:
  - Variable respuesta dicotómica tipo 0/1 (mujer sana/enferma).
  - Variables explicativas recogen la expresión diferencial de cada proteína.
- 4.3 Método de mínimos cuadrados parciales combinados con procedimientos de análisis discriminante (PLS-DA) para la reducción de la dimensionalidad de la matriz de diseño en relación al número de proteínas.

5. Resultados

- 5.1 Comparaciones múltiples a través del FDR.
  - Tasa de error $\alpha = 0.05$
  - Librerías multtest y genefilter de R.
  - 129 contrastes significativos antes de la corrección FDR.
  - Tres contrastes significativos después de la corrección FDR: Proteínas 1298, 1431 y 1672.
- 5.2 Modelos de regresión logística con penalización Lasso.
  - Aplicamos la penalización Lasso para reducir el número de proteínas del modelo eliminando aquellas que no están asociadas con la variable respuesta y obtenemos las proteínas: 798, 1298, 1431 y 1672.



  1. Muestras clasificadas correctamente según tejidos sanos o enfermos.
  2. Proteínas clasificadas en dos grupos según el valor del coeficiente estimado: uno formado por la proteína 798 con estimación del coeficiente negativo y otro formado por el resto de proteínas con estimación de coeficiente positivo.
- 5.3 Método PLS-DA de reducción de la dimensionalidad.
  - Librería mixOmics de R.
  - La función vip score mide la importancia de cada una de las variables explicativas para predecir la variable respuesta.
  - Seleccionamos las proteínas asociadas a las variables predictoras con vip$>$ 1.5 por ser aquellas variables que tienen una *alta influencia* en la variable respuesta.
  - Vip score de las proteínas seleccionadas son: 86, 655, 1328 y 1925.
  - Buscamos el valor vip de las proteínas seleccionadas con el modelo de regresión Lasso y vemos que tienen un valor de 1.4, luego tienen influencia en la variable respuesta aunque no una alta influencia.
  - La diferencia de los resultados podría deberse a la colinealidad de los datos.
  - Este método hace una buena clasificación de las muestras de tejido.



  1. Se diferencian claramente los tejidos sanos de los enfermos.
  2. Dentro de los tejidos enfermos se diferencian el tejido tipo CTA de los tejidos Micro y Macro.

6. Conclusiones

- Aplicamos contrastes múltiples con la corrección FDR a nuestros datos obteniendo 3 proteínas significativas.
- Lasso identifica cuatro proteínas siendo tres de ellas iguales a las obtenidas mediante contrastes múltiples.
- PLS-DA hace una buena clasificación de las muestras y con vip decidimos cuántas proteínas serán relevantes para explicar la variable respuesta.

Proyecto de investigación *Combination and Propagation of Uncertainties, COMPRO_UN*, MTM2013-42323-P, del Ministerio de Economía y Competitividad del Gobierno de España.

rgavidiaj@gmail.com, carmen.armero@uv.es, mluz.valero@uv.es, sandelpi@uv.es