XIX REUNIÓN CIENTÍFICA DEL GRUPO ARGENTINO DE BIOMETRÍA
SANTIAGO DEL ESTERO ARGENTINA • 2014

CONTROL DE CALIDAD DE MICROARREGLOS DE ADN MEDIANTE DESCOMPOSICIÓN ANOVA-PCA/PLS

CRISTÓBAL FRESNO[1], GERMÁN A GONZALEZ[1], GABRIELA A MERINO[1], JUAN C RODRIGUEZ[1], MÓNICA G BALZARINI[2], ELMER A FERNÁNDEZ[1]

[1]CONICET, Universidad Católica de Córdoba, Argentina.
[2]Grupo de Biometría, Facultad de Agronomía, Universidad Nacional de Córdoba, Argentina.

cfresno@bdmg.com.ar

RESUMEN

El control de calidad es una etapa fundamental para remover artefactos técnicos en datos de expresión de genes obtenidos a través de microarreglos de ADN. Los abordajes tradicionales sólo emplean estrategias univariadas, con el fin de comprobar el supuesto de normalidad global de los genes, al igual que para obtener valores de expresión comparables entre las diferentes muestras. No obstante, los anteriores presentan dos falencias: i) no incluyen la información del diseño experimental en el control de calidad; ii) la exploración de los genes se realiza de forma univariada. En este contexto, se presenta como alternativa para el control de calidad, descomponer los valores observados de expresión utilizando la información del diseño experimental a través de un análisis de la varianza (ANOVA). Luego, la contribución de los diferentes factores puede ser analizada de forma multivariada mediante dos técnicas de exploración como los son el Análisis de Componentes Principales (PCA por sus siglas en inglés) y la regresión de Mínimos Cuadrados Parciales (PLS por sus siglas en inglés). La estrategia propuesta ha sido aplicada con éxito sobre un experimento de microarreglos de dos factores. Con ella se ha logrado detectar y remover artefactos no considerados en el diseño experimental, que no pudieron ser detectados ni removidos por los abordajes tradicionales.

Palabras clave: ***exploración multivariada***, ***modelos lineales, diseño de experimentos***.

Introducción.

Las tecnologías de alto rendimiento, como los microarreglos de ADN, permiten explorar los valores de expresión de la totalidad de los genes de un organismo de forma simultánea, a partir de la lectura de intensidades de imágenes bidimensionales [1]. Esta tecnología se utiliza desde hace más de 20 años. Pese a esto, la comunidad científica ha reconocido la existencia de diferentes sesgos que tornan compleja la comparación entre muestras (variabilidad entre modelos de microarreglos, escáner utilizado, técnico que realizó el experimento, etc.) [1].

El control de calidad de los datos se realiza en gran medida a través de una exploración gráfica de los mismos con técnicas como gráficos de caja, de densidad y de dispersión. También se han diseñado métodos gráficos y analíticos específicos para ciertas plataformas de microarreglos, que hacen uso de las características propias de las mismas. En algunos casos es necesario aplicar transformaciones sobre los datos, a los efectos de obtener valores de expresión comparables entre las diferentes muestras y permitir satisfacer los supuestos de los modelos a aplicar en etapas posteriores. Sin embargo, no se cuenta actualmente con métodos que incorporen la información del diseño experimental para explorar la existencia de patrones de interacción multidimensionales entre factores no considerados en el experimento.
En el presente trabajo se propone realizar el control de calidad de datos genómicos, mediante una exploración multivariada PCA/PLS sobre la contribución de expresión de los diferentes factores del diseño, obtenidos a través de una descomposición ANOVA.

Desarrollo.

La metodología propuesta consiste en aplicar una descomposición ANOVA sobre los valores de expresión obtenidos con microarreglos de ADN. Esto consiste en expresar la matriz de expresión (genes en filas y muestras en columnas), como la suma de la contribución de una matriz de media global, las matrices asociadas a los diferentes factores (tiempo, concentración e interacción entre tiempo y concentración) y un término de error. Para ello se ajusta de forma secuencial, un modelo lineal para obtener la contribución de cada factor correspondiente para cada gen. Una vez obtenidas las matrices de cada efecto y residuos de cada ajuste, es posible realizar una exploración multivariada mediante PCA y/o PLS, lo que se conoce como ANOVA-PCA [2] y ANOVA-PLS [3] respectivamente. Estas metodologías se encuentran disponibles en el paquete *lmdme* de R [3] en el repositorio Bioconductor (www.bioconductor.org).

El control de calidad propuesto fue aplicado en un experimento en el cual se midió la expresión de genes en líneas celulares de melanoma, en tres tiempos diferentes bajo tres concentraciones de proteínas en el medio de cultivo [3]. Los dueños de los datos poseen evidencia, proveniente de experimentos previos, que soporta la existencia de genes que presentan interacción de tiempo por concentración a un nivel de confianza del 95%. No obstante, tras realizar un análisis de expresión diferencial mediante modelos lineales, no se encontraron genes para los cuales fuera significativa esta interacción.

Utilizando el paquete *lmdme* se realizó una descomposición ANOVA sobre los valores de expresión. Sobre la matriz de residuos obtenida para el término de interacción (tiempo por concentración) se realizó un PCA, del cual se graficaron las primeras dos componentes en un biplot en busca de artefactos no esperables.

Resultados y Conclusiones.

La Figura 1 muestra los biplots que se obtienen de aplicar ANOVA-PCA sobre los residuos del término de interacción. En el panel (1a) se aprecia la presencia de un patrón que tiende a agrupar en tres grupos los diferentes niveles de la interacción entre tiempo y concentración. Mediante la consulta al grupo de investigación que generó los datos, se determinó que los reactivos usados en el experimento se recibieron en tres envíos diferentes lo cual se corresponde con el patrón observado. En la Figura (1b), se observa que los agrupamientos coinciden con el mes de obtención de los datos. Este factor no fue considerado en el diseño original y claramente impacta en la obtención de resultados. Una conclusión similar se obtuvo al aplicar ANOVA-PLS para estudiar la covarianza entre la fecha de obtención y la matriz de residuos como se aprecia en la Figura (1c).

Posteriormente se repitió el análisis de expresión diferencial incluyendo el efecto de la fecha de obtención de los datos. De esta manera, fue posible encontrar 13 genes que presentaron interacción tiempo por concentración, como lo indicaba la evidencia experimental previa. Adicionalmente se comprobó la remoción del patrón encontrado con anterioridad, como se muestra en el biplot del panel (1d).

En conclusión, la aplicación de una descomposición ANOVA-PCA/PLS ha mostrado ser de gran utilidad para realizar un control de calidad multivariado sobre valores de expresión de genes obtenidos con microarreglos de ADN. En la base de datos utilizada fue posible encontrar un patrón multidimensional anómalo, debido a un factor no considerado originalmente en el análisis. Este efecto fue exitosamente removido y permitió así obtener conclusiones biológicas similares a las sugeridas en experimentos previos.

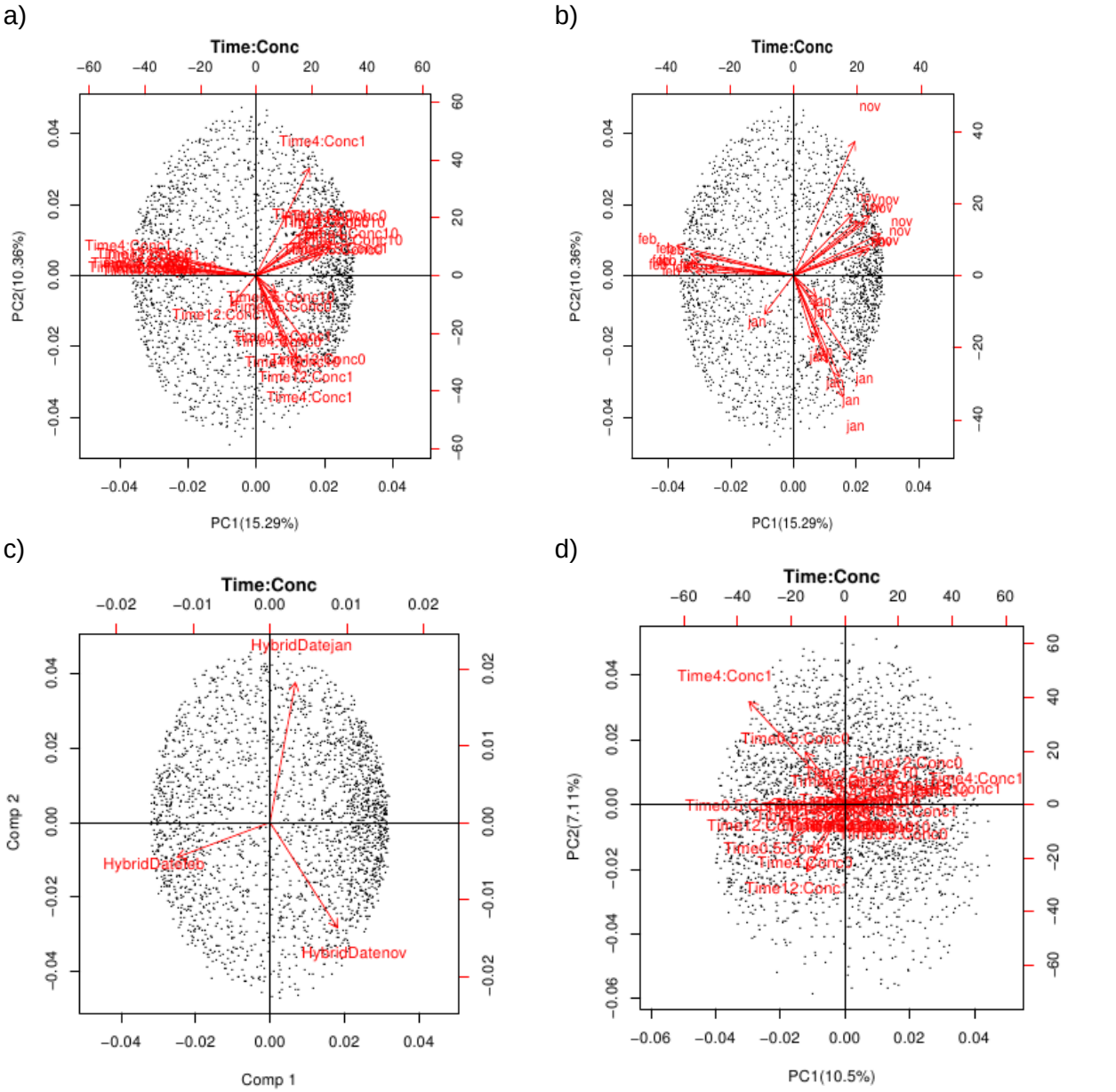

Figura 1: ANOVA-PCA sobre los datos de entrada del modelo de interacción. a) modelo original utilizando los niveles de la interacción entre tiempo y concentración. Se observan tres agrupamientos. b) modelo original con la etiqueta del mes de obtención de datos. Los agrupamientos corresponden con los meses de obtención de datos. c) ANOVA-PLS con la información de la fecha. Se comprueba el patrón observado con ANOVA-PCA. d) modelo incluyendo el efecto fecha, donde no se observan patrones (se removió el artefacto encontrado).

Bibliografía.

[1] Carey, V., et al. (2005) Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York.

[2] De Haan, J. et al. (2007). “Interpretation of ANOVA Models for Microarray Data Using PCA.” Bioinformatics, 23(2), 184-190.

[3] Fresno, C. et al (2014). “lmdme: Linear Model on Designed Multivariate Experiments in R”. Journal of Statistical Software, 56(7).