

¿EXISTE SIMETRÍA EN EL NIVEL DE EXPRESIÓN DEL TRANSCRIPTOMA?

Casellas, J.¹, Varona, L.
Genética i Millora Animal. IRTA-Lleida. 25198 Lleida.
joaquim.casellas@irta.es

INTRODUCCIÓN

Los recientes avances tecnológicos en el campo de la expresión génica proporcionan un marco de trabajo idóneo para determinar la base genética de los caracteres fenotípicos. No obstante, técnicas como los *microarrays* de cDNA, generan bases de datos enormes con un nivel de replicación mínima, lo cual requiere del desarrollo de métodos estadísticos apropiados para abordarlas. Los modelos mixtos han sido recientemente adaptados al análisis de *microarrays* (Wolfinger *et al.*, 2001), y con ellos la asunción de una distribución Gaussiana (simétrica) para los distintos efectos aleatorios del modelo se aplica sistemáticamente, sin cuestionarse su idoneidad en cada caso. Resulta difícil aceptar que los niveles de expresión génica de los diferentes transcritos de un tejido determinado puedan representar una distribución simétrica. En el mismo sentido, la simetría en la expresión diferencial de dos tejidos o estados metabólicos distintos no es más que una asunción simplista dentro de un amplio rango de posibilidades. En este sentido, el objetivo de este trabajo es el de modelar diferentes fuentes de asimetría en los modelos mixtos para análisis de *microarrays*.

MATERIAL Y MÉTODOS

Modelo de análisis

Tomando como punto de partida los datos generados a partir de *microarrays* de hibridación no competitiva, podemos asumir que un modelo básico de análisis sería:

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{d} + \mathbf{e}$$

donde \mathbf{y} es el vector de registros de expresión génica, \mathbf{X} es la matriz de incidencias para el efecto de cada *array* (\mathbf{a}), \mathbf{Z}_1 es la matriz de incidencias para el efecto de los distintos genes (\mathbf{g}), \mathbf{Z}_2 es la matriz de incidencias para los efectos de expresión diferencial jerarquizados a gen (\mathbf{d}), y \mathbf{e} es el vector de residuos. Siendo p el número de arrays, q el número de transcritos analizados, y asumiendo que los datos en \mathbf{y} están ordenados por gen dentro de *array*, podemos definir la verosimilitud Bayesiana como:

$$p(\mathbf{y}|\mathbf{a}, \mathbf{g}, \mathbf{d}, \mathbf{R}) \sim N(\mathbf{X}\mathbf{a} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{d}, \mathbf{I} \otimes \mathbf{R})$$

siendo \mathbf{I} una matriz identidad de dimensiones $p \times p$, y \mathbf{R} una matriz $q \times q$ con ceros fuera de la diagonal y las varianzas residuales para cada gen en los elementos diagonales (modelo de varianzas residuales heterogéneas). Podemos asumir distribuciones *a priori* uniformes para \mathbf{a} y \mathbf{R} , mientras que \mathbf{g} y \mathbf{d} representan dos fuentes potenciales de asimetría distribuidas según:

$$p(\delta | \sigma_\delta^2, \lambda_\delta) \propto \prod_{i=1}^q \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left\{-\frac{(\delta_i - \lambda_\delta x_{\delta,i})^2}{2\sigma_\delta^2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x_{\delta,i}^2}{2}\right\} dx \quad \delta = \{\mathbf{g}, \mathbf{d}\}$$

donde λ_δ es el parámetro de asimetría con distribución *a priori* plana, y $x_{\delta,i}$ es un parámetro auxiliar procedente de una distribución normal estándar truncada en 0 ($0 \leq x_{\delta,i} < \infty$; Sahu *et al.*, 2003). La inferencia sobre los distintos parámetros del modelo se efectuará a partir de muestreos iterativos de Gibbs estándares (Wang *et al.*, 1994), con la salvedad de $x_{\delta,i}$ que será un valor positivo procedente de una distribución normal truncada en 0.

¹ Trabajo realizado en el marco del programa “Juan de la Cierva” del Ministerio de Educación y Ciencia.

Ejemplo

El modelo descrito anteriormente se ha aplicado sobre datos de expresión génica en fibroblastos de *Homo sapiens* (Hombre, $n = 18$), *Gorilla gorilla* (Gorila, $n = 11$) y *Pan paniscus* (Chimpancé, $n = 10$), obtenidos mediante *microarrays* no competitivos (*Human Genome U95 Set*, Affymetrix). La obtención de este material experimental, procesado, y resultados del análisis mediante metodología estadística estándar pueden consultarse en Karaman *et al.* (2003). Los datos completos de expresión génica para cada *array* son de dominio público y pueden obtenerse en *Gene Expresión Omnibus*² (referencia GDS340). Después de descartar los *loci* con nivel nulo de expresión, el análisis se efectuó sobre el logaritmo neperiano de la expresión de 3.700 transcritos. Se realizaron tres análisis paralelos para determinar la expresión diferencial entre *Homo sapiens* y *Gorilla gorilla*, *Homo sapiens* y *Pan paniscus*, y *Gorilla gorilla* y *Pan paniscus* así como el grado de asimetría en la distribución de efectos aleatorios del modelo. Para cada análisis se lanzó una única cadena de 500.000 elementos, descartando los primeros 50.000 como *burn-in* (Raftery y Lewis, 1992).

RESULTADOS Y DISCUSIÓN

La implementación de efectos aleatorios con distribuciones asimétricas en el análisis de *microarrays* implica un avance metodológico substancial a la hora de modelar los registros de expresión génica. Aunque la asunción típica de distribuciones normales simétricas para ese tipo de efectos se fundamenta en el compromiso entre plausibilidad biológica y simplicidad computacional, la inclusión de asimetría y el posterior análisis de los parámetros adicionales, resulta trivial, tal como se describe en el apartado de Material y Métodos.

El parámetro λ caracteriza el grado de asimetría de la distribución, convergiendo a una distribución normal de media cero cuando $\lambda = 0$. Aunque la asimetría caracterizada por λ depende también de la varianza del efecto (Sahu *et al.*, 2003), un valor de λ positivo incrementa la probabilidad a la derecha del valor modal, mientras que una λ negativa tiene el efecto inverso. Los análisis efectuados revelaron un grado substancial de asimetría positiva en la expresión génica global (Tabla 1, Figura 1a), muy semejante en los tres análisis, resultado que confirma nuestra hipótesis de partida en cuanto a la asimetría de la expresión del transcriptoma.

Tabla 1. Estimación del parámetro de asimetría para la expresión génica basal (λ_g) y la expresión diferencial (λ_t) en los distintos análisis.

	λ_g		λ_t	
	Moda	HPD95	Moda	HPD95
<i>Homo sapiens</i> vs. <i>Gorilla gorilla</i>	1,65	1,61 a 1,70	0,47	0,44 a 0,50
<i>Homo sapiens</i> vs. <i>Pan paniscus</i>	1,65	1,60 a 1,70	0,44	0,42 a 0,47
<i>Gorilla gorilla</i> vs. <i>Pan paniscus</i>	1,64	1,58 a 1,68	-0,33	-0,37 a -0,28

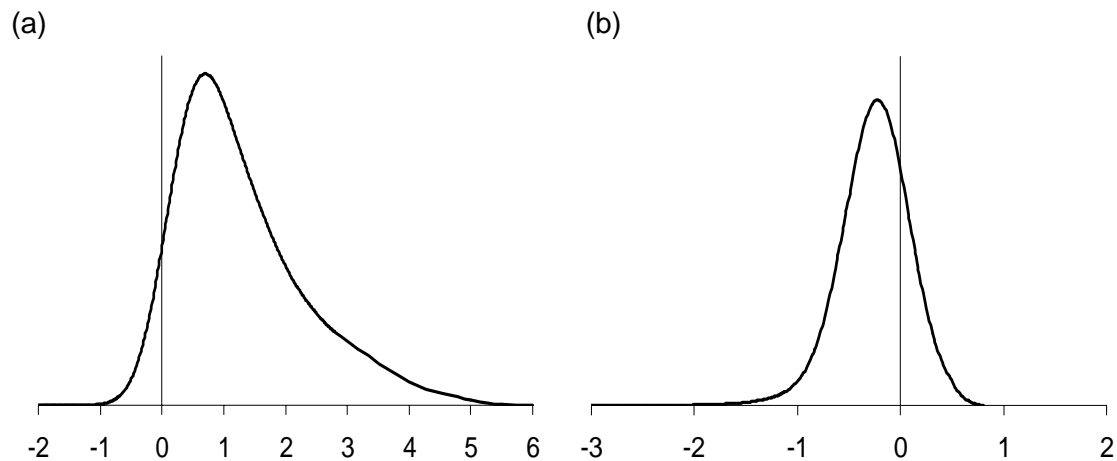
HPD95: Región de máxima probabilidad posterior al 95 %.

La distribución de la expresión diferencial en los distintos tejidos también fue asimétrica, aunque de manera mucho más moderada (Figura 1b). No obstante, resulta importante destacar que este ejemplo se centró en el mismo tejido aunque de especies distintas, lo cual origina una variabilidad moderada de la expresión génica diferencial (Figura 1b). En este sentido sería esperable obtener desviaciones más importantes de la simetría en comparaciones entre tejidos más diferenciados, como sería el ejemplo clásico del análisis entre células sanas y células cancerígenas. Aunque el análisis conjunto de miles de genes aumenta la probabilidad de obtener una expresión diferencial equilibrada entre las dos

² <http://www.ncbi.nlm.nih.gov/geo/>

especies analizadas, las desviaciones observadas sugieren que la sobre-expresión de algunos genes en una especie no se ve necesariamente compensada por la sobre-expresión del mismo número de genes en la otra, al menos no con las mismas magnitudes de expresión génica. Resulta curioso destacar que dada la asimetría de la distribución, existe más expresión génica diferencial a nivel de fibroblastos en *Gorilla gorilla* o *Pan paniscus* que en *Homo sapiens*, y en caso de comparar ambas especies de primates africanos, la distribución se desvía a favor del gorila.

Figura 1. Distribución de las estimas obtenidas (media posterior) para los efectos aleatorios de gen (a) y tratamiento dentro de gen (b) en el análisis *Gorilla gorilla* vs. *Pan paniscus*.



REFERENCIAS BIBLIOGRÁFICAS

- Karaman, M. W., Houck, M. L., Chemnick, L. G., Nagpal, S., Chawannakul, D., Sudano, D., Pike, B. L., Ho, V. V., Ryder, O. A., Hacia, J. G. 2001. Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. *Genome Res.*, 13, 1619-1630.
- Raftery, A. E., Lewis, S. M. 1992. How many iterations in the Gibbs sampler? Páginas 763-774 en *Bayesian Statistics IV* (Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M.), Oxford University Press, NY.
- Sahu, S. K., Dey, D. K., Branco, M. D. 2003. A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian J. Stat.*, 31, 129-150.
- Wang, C. S., Rutledge, J. J., Gianola, D. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genet. Sel. Evol*, 26, 91-115.
- Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, R., Afshari, C., Paules, R. S. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, 8, 625-637.