Métodos robustos de normalización de microarreglos de ADNc basados en la mediana

Median based robust normalization methods of cDNA microarray data

Paredes, José *, Ramírez, Juan
Postgrado de Ing. Biomédica, Facultad de Ingeniería, ULA,
Mérida 5101, Venezuela
*paredesj@ula.ve

Bianchi, Giorgio
Departamento de Matemática, Facultad de Ciencias, ULA
Mérida 5101, Venezuela

Recibido: 25-11-2005 Revisado: 17-05-2006

Resumen

En este trabajo se presentan dos nuevos métodos de normalización de datos microarreglos de ADNc basados en el operador de mediana ponderada. El primer método estima el parámetro de normalización usando regresión por desviación absoluta mínima. Tal método asume que las variaciones entre los datos generados por repeticiones del mismo experimento tienen características impulsivas, y por tanto pueden ser modeladas por distribuciones de tipo Laplaciano. El segundo método incorpora robustez a un procedimiento de normalización previamente reportado sustituyendo la correlación tradicional por la correlación basada en mediana. El desempeño de los dos métodos propuestos es comparado con métodos similares reportados en la literatura. Las medidas de desempeño usadas son el error medio cuadrático y el error medio absoluto entre el conjunto de datos de referencia y el conjunto de datos normalizados. Además se evalúa la variación de los datos normalizados usando gráficos de cuartiles.

Palabras Claves: Microarreglos, normalización, mediana ponderada.

Abstract

This paper introduces two new methods to normalizing microarray expression data based on weighted median. The first approach exploits the fact that variations between replicated slides of the same experiment have impulsive characteristic and, therefore, they are better modeled by a Laplacian distribution leading to a least absolute deviation regression method for the estimation for the scaling parameter. The second approach adds robustness to a previously reported method derived using linear regression by replacing traditional correlation by a Median based correlation. The performances of the proposed methods are compared to those yielded by well-known methods reported in the literature using, as performance measure, the mean square error and the mean absolute error between the reference data set and the normalized set. Furthermore, variation of the normalized data is evaluated using boxplots

Key words: Microarray, normalization, weighted median.

1 Introducción

Los microarreglos de ADN complementario (ADNc) constituyen un nuevo tipo de tecnología genética que permite el monitoreo de los niveles de expresión de miles de genes simultáneamente. Las aplicaciones de este conjunto de

técnicas van desde el análisis de los niveles de expresión genética de los organismos en distintas condiciones ambientales hasta la caracterización de la expresión genética de tumores procedentes de pacientes con cáncer. Esta diversidad de aplicaciones ha facilitado el desarrollo en la investigación clínica y farmacéutica, obteniéndose diagnósticos más acertados de enfermedades y estudios más detallados de la efectividad de los tratamientos (Choi, 2004; Yang y col., 2004).

Actualmente, uno de los problemas presentes en la tecnología de microarreglos de ADNc es la poca reproducibilidad del experimento, debido, fundamentalmente, a la influencia de diversas fuentes de error en la medición de los niveles de expresión genética. Tales fuentes de error son atribuidas, principalmente, a la cantidad de muestra usada en cada experimento, a la cantidad y efectividad del tinte aplicado a las muestras y a las variaciones en los parámetros del escáner (Wang y col., 2002, Schuchhardt y col. 2000). La realización de múltiples repeticiones del mismo experimento de ADNc ha sido uno de los mecanismos usados con el fin de obtener valores de niveles de expresión genética más confiables a partir de la imágenes en niveles de grises obtenidas del experimento (Lee y col., 2000). Sin embargo, repeticiones del mismo experimento de microarrreglos no producen los mismos resultados de niveles de expresión genética, impidiendo un análisis correcto de los resultados. Con el objeto de remover o minimizar las variaciones de los niveles de expresión genética en repeticiones del mismo experimento de microarreglo, ha surgido la normalización entre múltiples repeticiones como una etapa necesaria de preprocesamiento de los datos, antes del análisis genético de los mismos.

Uno de los criterios usados en la normalización de datos de microarreglos procedentes de múltiples repeticiones, asume que la relación entre dos conjuntos de datos está definida por un factor de escala. La estimación del factor de escala (parámetro de normalización) usualmente asume que la distribución de los errores entre un par de conjuntos de datos siguen un modelo gaussiano (Chen y col., 1997). Sin embargo, se ha reportado en (Purdom y Holmes, 2005) que los datos generados por los experimentos de microarreglos obedecen a modelos no Gaussianos con características impulsivas, las cuales pueden ser mejor modelados a través de distribuciones Laplacianas. Como consecuencia, los métodos de normalización derivados bajo la suposición Gaussiana tienen un pobre desempeño en presencia de datos de naturaleza impulsiva.

En este trabajo se proponen dos nuevos métodos de normalización de múltiples repeticiones de datos de microarreglos de ADNc que se derivan como consecuencia de la suposición que las diferencias entre múltiples repeticiones del mismo experimento siguen un modelo Laplaciano conllevando al uso del operador de mediana ponderada en el proceso de normalización. El primer método emplea regresión lineal basada en desviación absoluta mínima (least absolute deviation) para el cálculo del factor de escala. Este método explota el hecho de que las variaciones entre dos conjuntos de datos de microarreglos siguen modelos estadísticos de colas pesadas, similares a los presentados por las distribuciones Laplacianas. La segunda propuesta de normalización añade robustez en la estimación del factor de escala, sustituyendo el operador de correlación lineal previamente reportado en (Zar, 1999), por el operador de correlación basado en mediana. Ambos métodos presentan un buen desempeño en presencia de impulsos debido a las características robustas implícitas en el operador de mediana ponderada.

El desempeño de los métodos de normalización propuestos se cuantifica con el cálculo del error medio cuadrático y el error medio absoluto existente entre un par de conjuntos de datos de microarreglos y se comparan con los métodos previamente reportados en (Zar, 1999, Chen y col., 1997, Hedge y col. 2000).

La organización del presente artículo es la siguiente. En la sección 2, se describe brevemente el experimento de microarreglos de ADNc. Seguidamente, en la sección 3 se explican los métodos de normalización que estiman el factor de escala entre un par de conjuntos de datos, usando la teoría lineal. En la sección 4 y en la sección 5 se describen respectivamente, la propuesta de normalización basada en regresión lineal robusta y el método de normalización que emplea la correlación basado en mediana. En la sección 6, se muestran los resultados de la implementación de los métodos propuestos y se analizan los resultados. Finalmente, en la sección 7, se exponen las conclusiones generadas del presente trabajo.

2 Breve descripción del experimento de microarray

El experimento de microarreglos de ADNc es un conjunto de técnicas estrictamente diseñadas con el objeto de obtener los valores de los niveles de expresión genética de ciertos genes en las células en estudio. Tal experimento, se realiza sobre una lámina de vidrio que contiene miles de pequeños envases llamados spots, y dentro de cada uno de los spots, están contenidas cadenas conocidas de ADNc (genes conocidos). Paralelamente, se extraen las cadenas de ácido ribonucleico mensajero (ARNm) a dos tejidos que serán objeto de análisis, en donde las cadenas obtenidas de uno de los tejidos se tomarán como muestras de referencia y las cadenas procedentes del otro tejido serán las muestras de prueba.

Por ejemplo, en el estudio de los niveles de expresión genética de células cancerosas, se extraen las cadenas de ARNm de las células procedentes de tejido sano (muestras de referencia), y de células de tejidos con cáncer (muestras de prueba). Luego, a las cadenas de ARNm obtenidas de cada tejido, se etiqueta con tintes de color distintos. Los colores empleados en el etiquetado son el tinte verde y el tinte rojo.

Una vez etiquetadas las muestras son vertidas en cada uno de los spots con el objeto de estimular un proceso de

hibridación genética entre las cadenas de ARNm etiquetadas y las cadenas de ADNc conocidas y contenidas en los spots. En la Fig. 1 se despliega de forma gráfica los principales pasos del experimento de microarreglos de ADNc.

Fig. 1. .Pasos empleados en el experimento de microarreglos de ADNc.

Luego de la hibridación, se pasa el arreglo a través de un escáner que excita los tintes aplicados a las muestras. El escáner primero aplica un láser con una longitud de onda que excita los tintes verdes, y luego se excita el arreglo con el láser que estimula los tintes rojos. La aplicación del escáner sobre el arreglo genera dos imágenes, una en verde y una en rojo. La imagen verde es llamada también imagen de canal-3, y la imagen roja es denominada imagen de canal-5. Cada imagen obtenida es un archivo en formato TIFF que tiene la capacidad de desplegar la información de aproximadamente 43.000 *spots* a 2^{16} = 65536 niveles de intensidades distintos. En la Fig. 2 se muestra gráficamente el proceso de obtención de las imágenes de microarreglos.

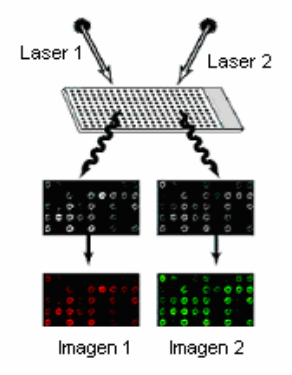

Fig. 1: Proceso de extracción de las imágenes de microarreglo.

Los valores de los niveles de expresión genética se obtienen a partir del par de imágenes a niveles de gris obtenidas. La imagen de Canal-3 contiene los niveles de expresión genética de las muestras etiquetadas con tinte verde, la cual ha sido nuevamente representada en la Fig. 2 como la Imagen 2 a niveles de verde y en la imagen de Canal-5 están los niveles de expresión genética de las muestras etiquetadas con el tinte rojo, mostrada como la imagen 1 en la Fig. 2. Con el objeto de obtener los niveles de expresión, se aplica cada imagen un conjunto de técnicas de tratamiento de imágenes. Primero, la imagen se enrejilla y se divide en tantas regiones como *spots* contenga el microarreglo. Luego, se procede con la segmentación de la imagen del *spot*, la cual divide cada rejilla en dos regiones, una región perteneciente al *spot* llamada *foreground*, y la región que forma parte del fondo o *background*. Así, el nivel de expresión genética del *spot* en cada rejilla se define como:

$$NEG = mediana\{f_1, f_2, f_3, \ldots\} - mediana\{b_1, b_2, b_3, \ldots\}$$

donde *NEG* es el nivel de expresión genética del *spot* en estudio, $\{f_1, f_2, f_3, \ldots\}$ es el conjunto de píxeles del *foreground* y $\{b_1, b_2, b_3, \ldots\}$ es el vector de píxeles del *background*. Como puede observarse de la ecuación anterior se ha realizado una corrección de *background* a los niveles de expresión existentes en el *spot*. Es importante destacar que se estiman dos niveles de expresión genética por cada spot, un nivel de expresión obtenido del canal-5 y un nivel de expresión resultante del canal-3.

3 Métodos de normalización por escalamiento de repeticiones de microarrays.

Sea $\{Y_i\}$ y $\{X_i\}$ los niveles de expresión genética obtenidos a partir de dos repeticiones de microarray del mismo experimento, donde $i = 1,2,\ldots,N$ y N el número de genes en el microarray. Como se mencionó en la introducción, debido a los errores en el proceso de adquisición, los niveles de expresión en ambas repeticiones no son idénticos (Schuchhardt y col. 2000) por el contrario se originan fluctuaciones que deben ser corregidas mediante un proceso de normalización. Estas fluctuaciones, entre ambos conjuntos de datos, han sido apropiadamente modeladas como una relación lineal que involucra operaciones de escalamiento y desplazamiento (Wang y col. 2002), es decir:

$$Y_i = aX_i + b + \eta \qquad (1)$$

donde *a* y *b* son los parámetros de escala y de desplazamiento, respectivamente, y η representa los errores entre ambas repeticiones. En la Ec. (1) se ha considerado el conjunto $\{Y_i\}$ como el conjunto de datos de referencia y los $\{X_i\}$ como los datos a ser normalizados.

El proceso de normalización consiste en una transformación del conjunto de datos $\{X_i\}$ en un nuevo conjunto de datos, $\{X_i'\}$, mas confiable para su análisis. Dado que se ha asumido una relación lineal entre los pares de datos $\{Y_i, X_i\}$, el proceso de normalización puede lograrse en forma óptima mediante una transformación de carácter li-

neal, tal como se muestra a continuación:

$$X_i' = aX_i + b \tag{2}$$

donde *a* y *b* deben ser óptimamente obtenido usando los datos de referencia y los datos a normalizar.

A fin de mantener la simplicidad en el proceso de normalización, en este artículo se asume que el desplazamiento de los niveles de expresión genética medidos es igual a cero (*b*=0), por lo que la expresión genética de la repetición *Y* es una versión escalada de la repetición *X*. Esta suposición sigue el modelo planteado en (Chen y col., 1997; Hedge y col. 2000; Zar, 1999) y origina a su vez tres métodos de normalización.

3.1 Normalización basada en regresión lineal

Un primer método de normalización usa regresión lineal para determinar el parámetro de escalamiento (Zar, 1999). La idea fundamentalmente consiste en determinar el valor de a que minimice el error medio cuadrático entre el conjunto de datos de referencia y los datos normalizados. Es decir,

$$\tilde{a} = \arg\min_a \frac{1}{N}\sum_{i=1}^{N}[Y_i - aX_i]^2 \tag{3}$$

Así, el nuevo conjunto de datos normalizados viene dado por

$$X_i' = \tilde{a}X_i = \left(\frac{\sum_{i=1}^{N} X_i Y_i}{\sum_{i=1}^{N} X_i^2}\right) X_i \tag{4}$$

3.2 Normalización de media de razón unitaria

Un segundo método de normalización asume que el promedio de la razón de cada gen de referencia y cada gen a normalizar debe ser igual a uno (Chen y col., 1997). Bajo esta suposición el parámetro de normalización viene dado por:

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i} \tag{5}$$

3.3 Normalización de media unitaria

En este método de normalización no se requiere el cálculo explícito de un parámetro de escalamiento dado que cada repetición es normalizada independientemente de forma tal que la media global de todos los datos en cada repetición sea igual a uno. Es decir, si $\{X_i, Y_i\}$ son los pares de datos sin normalizar, el conjunto de datos normalizados son $\left\{X_i \Big/ \sum_{i=1}^{N} X_i , Y_i \Big/ \sum_{i=1}^{N} Y_i\right\}$. Obsérvese que multiplicando cada par de datos por el término $\sum_{i=1}^{N} Y_i$ se producen los datos normalizados donde el parámetro de normalización es dado por (Hedge y col. 2000)

$$\bar{a} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} \tag{6}$$

Las Ecs. (4), (5) y (6) producen, en general, diferentes tipos de normalización y todas ellas se originan bajo la suposición que el parámetro de desplazamiento es igual a cero.

4 Método de normalización basado en regresión robusta

El primer método de normalización propuesto en este artículo está basado en el hecho de que el tipo de errores generados durante el proceso de adquisición -- y que refleja variabilidad de la expresión genética entre repeticiones del mismo experimento -- puede ser modelado a través de una distribución de colas mas pesadas que la distribución Gaussiana (Bloch y Arce, 2002). Específicamente, en este artículo se propone modelar dicho error a través de una distribución Laplaciana. Esto conlleva naturalmente a que en lugar de utilizar el error medio cuadrático como criterio de optimización en el cálculo del parámetro de escalamiento, se utiliza el error medio absoluto.

El error medio absoluto, definido en términos de los niveles de expresión genética, entre los datos de referencia y los datos normalizados viene dado por

$$\varepsilon = \frac{1}{N}\sum_{i-1}^{N}\left|Y_i - aX_i\right| \tag{7}$$

La minimización del error medio absoluto (7) en función del parámetro de escalamiento *a* conduce a la determinación de dicho parámetro. La expresión (7) puede rescribirse como:

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N}\left|X_i\right|\left|\frac{Y_i}{X_i} - a\right| \tag{8}$$

Puede observarse en la Ec. (10) que $|X_i|$ hace las ve-

ces de ponderación, y la influencia de la muestra $\frac{Y_i}{X_i}$ en el cálculo del parámetro de escalamiento depende de la ponderación $|X_i|$. A la Ec. (10) se le conoce en la literatura como mediana ponderada (Arce 1998, Astola y col. 1997).

$$\hat{a} = \arg\min_a \frac{1}{N}\sum_{i=1}^{N}|X_i|\left|\frac{Y_i}{X_i} - a\right| \quad (9)$$

$$\hat{a} = MED\left(|X_1| \Diamond \frac{Y_1}{X_1}, |X_2| \Diamond \frac{Y_2}{X_2}, \ldots, |X_N| \Diamond \frac{Y_N}{X_N}\right) \quad (10)$$

donde MED es el operador mediana (no lineal) y el símbolo ◊ denota la operación de repetición, es decir

$$W \Diamond Z = \overbrace{Z, Z, Z, \ldots, Z}^{W \text{ veces}}$$

Así, siguiendo la distribución Laplaciana para modelar las fluctuaciones entre repeticiones la determinación del parámetro de normalización se reduce a repetir la muestra $\frac{Y_i}{X_i}$, $|X_i|$ veces, ordenar el conjunto de muestras repetidas y seleccionar la muestra del medio como parámetro de escalamiento.

Sin embargo, determinar la mediana ponderada usando este algoritmo es computacionalmente ineficiente dado que el ordenamiento se realiza sobre el conjunto ampliado de muestras repetidas, cuyo número de veces de repetición depende de los niveles de expresión genética del conjunto X, el cual a su vez puede tomar valores entre 0 y 216-1. Por consiguiente, se hace necesario utilizar una definición más general del operador de repetición ◊ que permita incluso el uso de ponderaciones reales no negativas.

En (Astola y col. 1997) se presenta esta extensión, por lo que el parámetro de escalamiento, cuando la expresión genética $|X_i|$, es un número real cualquiera se determina como sigue:

- Calcular el valor umbral definido como $T_u = \frac{\sum_{i=1}^{N}|X_i|}{2}$.
- Ordenar el conjunto de muestras $\frac{Y_i}{X_i}$ de mayor a menor.
- Sumar las ponderaciones correspondientes a la muestras ordenadas comenzando desde la mayor y siguiendo en orden hasta la menor.
- Seleccionar como parámetro de escalamiento aquella muestra $\frac{Y_i}{X_i}$ cuya ponderación hace que la suma parcial de ponderaciones supere el valor umbral.

Es importante notar que usando este algoritmo sólo se ordenan N muestras y no $\sum_{i=1}^{N}|X_i|$ muestras. A fin de ilustrar este procedimiento, considere que del experimento de adquisición X e Y se tienen, respectivamente, el siguiente conjunto de expresiones genéticas {120, 200, 300, 250, 100} y {200, 500, 700, 400, 180}. En este caso el valor umbral es: T_u = 485. El conjunto de muestras $\frac{Y_i}{X_i}$, sus ponderaciones $|X_i|$, las muestras ordenadas y sus correspondientes ponderaciones, al igual que la suma parcial de ponderaciones, se muestra en la Tabla 1. Como se puede observar el parámetro de escalamiento es 2.333 dado que al comenzar por la muestra mas grande, el umbral no se supera sino al sumar la ponderación asociada con la muestra 2.333.

Tabla 1: Cálculo de la mediana ponderada

Muestras $\frac{Y_i}{X_i}$ originales	1.666	2.500	2.333	1.600	1.800
Ponderaciones correspondientes	120	200	300	250	100
Muestras $\frac{Y_i}{X_i}$ ordenadas	1.600	1.666	1.800	2.333	2.500
Ponderaciones de la muestras ordenadas	250	120	100	300	200
Suma parcial de ponderaciones	970	720	600	500	200

Finalmente, cabe mencionar que es común encontrar en la literatura analogías entre el operador de mediana y el operador de media (Arce 1998, Mitra y col. 2001). En nuestro método de determinación del parámetro de escalamiento también se encuentra tal analogía al observar con detalle la expresión (5) la cual es la media no ponderada de la muestras $\frac{Y_i}{X_i}$, y la expresión (10) la cual es la mediana ponderada de las muestras $\frac{Y_i}{X_i}$ con ponderaciones $|X_i|$.

5 Método de normalización basado en correlación robusta.

Un segundo método de normalización surge al usar los conceptos de correlación robusta, recientemente introducidos por Arce y Li en (Arce y Li, 2002), en el cálculo del pa-

rámetro de normalización. Al examinar el método de normalización basado en regresión lineal puede notarse que el parámetro de normalización no es mas que la relación entre la correlación existente en los niveles de expresión genética de los experimentos *X* e *Y*, y la autocorrelación de los datos obtenidos en el experimento *X*. Sin embargo, la naturaleza impulsiva de estos datos, ocasionados por los múltiples errores mencionados en la introducción, produce una estimación pobre de la correlación (Rodríguez y col. 2005) y como consecuencia una estimación errada del parámetro de normalización.

Extendiendo los conceptos introducidos en (Arce y Li, 2002) para determinar la correlación y autocorrelación en el cálculo del parámetro de escala, emerge, naturalmente, un segundo método de normalización basado en el operador de mediana.

Este método surge al sustituir la correlación tradicional, utilizada para el cálculo del parámetro de normalización en el método de regresión lineal, por la correlación robusta introducida en (Arce y Li, 2002) y estudiada en detalle en (Rodriguez, 2005). Específicamente el valor del parámetro de normalización a usando la correlación basada en mediana es:

$$\breve{a} = \frac{\left(\frac{1}{N}\sum_{i=1}^{N}|Y_i|\right) \bullet MED\left(|Y_i| \lozenge \operatorname{sgn}(Y_i) X_i \big|_{i=1}^{N}\right)}{\left(\frac{1}{N}\sum_{i=1}^{N}|X_i|\right) \bullet MED\left(|X_i| \lozenge \operatorname{sgn}(X_i) X_i \big|_{i=1}^{N}\right)} \qquad (11)$$

Expresión que contiene en el numerador la Correlación Mediana Muestral de los datos experimentales *X* e *Y*, y en el denominador la Autocorrelación Mediana Muestral de *X*. El hecho de que aparezcan $\operatorname{sgn}(Y_i)$ y $\operatorname{sgn}(X_i)$ en esta expresión, se debe a que de esta forma se define una estructura de filtro de mediana ponderada que admite pesos negativos (Arce, 1998), lo cual da lugar a la definición de correlación robusta basada en mediana (Arce y Li, 2002). Siendo el cálculo de correlación mas robusto, conducirá a valores mas confiables del parámetro de escalamiento y por tanto a una mejor normalización.

6 Resultados y discusión

Para implementar los métodos de normalización propuestos se tomaron 12 repeticiones de un mismo experimento de microarreglos. La cantidad inicial de genes, es decir el tamaño de cada secuencia $\{X_i\}$, fue de 18432. Debido al hecho de que en las ecuaciones (5) y (10) aparece X_i en el denominador de un cociente, se eliminaron en todas las repeticiones los genes para los cuales al menos uno de los valores de expresión genética era cero, quedando finalmente 18162 valores.

Teniendo 12 repeticiones de un mismo experimento, cualquiera de esos 12 conjuntos de datos podría fungir como conjunto de referencia para normalizar las otras 11 repeticiones con respecto a la repetición escogida. El proceso de evaluación del desempeño de los 5 métodos de normalización, el cual se describe en esta sección, se realizó 12 veces, una vez por cada conjunto referencial. Es decir, se consideraron 12 posibilidades distintas al tomar cada repetición como conjunto de referencia para la normalización de los restantes. Se observó que los resultados se mantenían consistentes a lo largo de cada una de las 12 posibilidades, y por lo tanto se procedió a escoger una de ellas que fuera representativa de los resultados generales.

En definitiva, lo que se presenta a continuación corresponde a tomar la primera realización del experimento como referencia de normalización. Se procedió entonces a normalizar las restantes 11 repeticiones usando los 5 métodos descritos previamente, obteniéndose así 5 veces 11 parámetros de normalización con los cuales se hallaron nuevos conjuntos de datos normalizados.

Para evaluar el desempeño de los diversos métodos se compararon los errores medios cuadráticos y los errores medios absolutos generados por los 5 métodos para cada repetición del experimento.

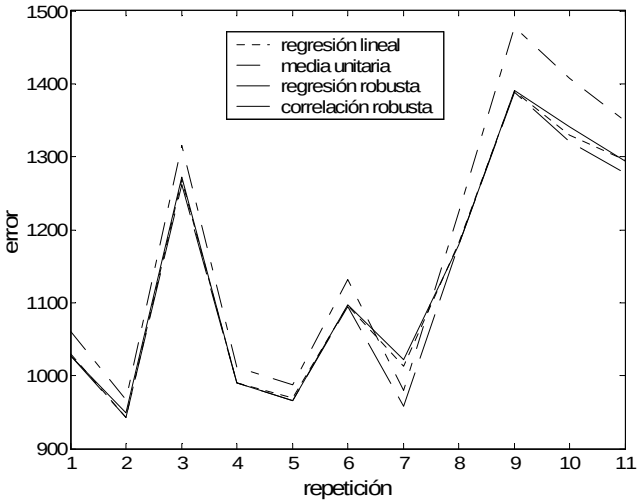

Fig. 3: Error medio absoluto: ——: regresión robusta, ····: regresión lineal, ———: correlación robusta, —·— : media unitaria.

Las figuras 3 y 4 muestran, respectivamente, el error medio absoluto y el error medio cuadrático para *cuatro* de los métodos de normalización, a saber: normalización basada en regresión lineal (línea punteada y símbolo △), normalización de media unitaria (línea a puntos y trazos), normalización basada en regresión robusta (línea a trazos y símbolo ○), y normalización basada en correlación robusta (línea sólida y símbolo *). El eje de las abscisas representa las distintas repeticiones del experimento, mientras que el eje de las ordenadas representa el valor del error.

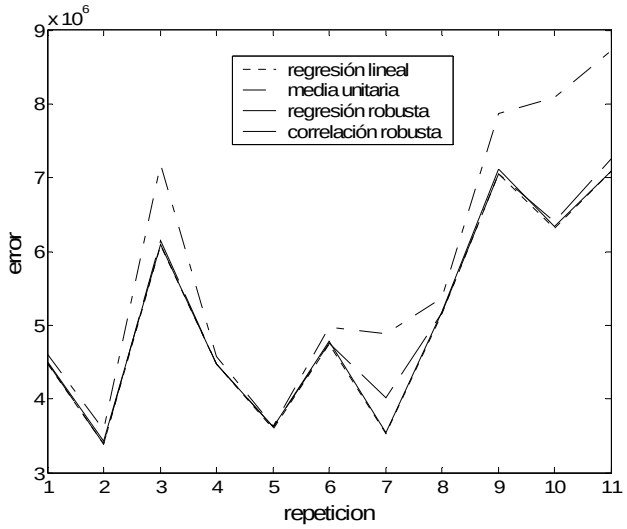

Fig. 4: Error medio cuadrático: ——: regresión robusta, ····: regresión lineal, ————: correlación robusta, —·— : media unitaria.

El método de normalización de media de razón unitaria no ha sido representado en ninguna de las dos figuras debido a que los errores generados por este método son de un orden de magnitud superior que los representados en las gráficas (como puede apreciarse en las Tablas 2 y 3), y por lo tanto, para evitar pérdida de resolución a lo largo del eje vertical, no ha sido graficado.

Como es de esperarse, tanto el método de regresión lineal como el método de regresión robusta, al estar diseñados para minimizar respectivamente el error medio cuadrático y el error medio absoluto, se comportan de manera óptima de acuerdo a la medida de desempeño usada. Sin embargo, como puede observarse en la Fig. 4, el método de regresión robusta a pesar de no ser el más eficiente en el sentido del error medio cuadrático, se mantiene en un nivel competitivo comparable con los métodos tradicionales basados en regresión lineal y media unitaria. El método basado en correlación robusta es un poco más equilibrado en este sentido, ya que tanto en una como en otra medida de desempeño se mantiene prácticamente a la par del método tradicional basado en regresión lineal.

Puede observarse también que la normalización de media unitaria es el menos eficiente de los cuatro métodos representados gráficamente.

Las Tablas 2 y 3 muestran numéricamente la misma información representada gráficamente en las figuras 3 y 4 (además de incluir el método de media de razón unitaria), para un análisis más preciso. Como puede observarse de los errores medios cuadráticos y errores medios absolutos, el método de media de razón unitaria (Raz uni) presenta un pobre desempeño.

Tabla 2: Error medio absoluto

	Reg lin	Raz uni	Med uni	Reg rob	Corr rob
1	1028.9	12374.0	1059.6	1026.5	1026.5
2	942.8	4297.2	966.7	942.7	949.2
3	1268.0	4950.1	1315.5	1261.5	1272.4
4	990.2	2922.9	1011.6	989.6	989.7
5	970.1	5754.9	987.5	965.4	965.4
6	1095.8	6533.8	1132.2	1095.1	1097.2
7	1013.2	1367.3	979.5	957.7	1021.4
8	1181.3	1511.2	1224.1	1178.7	1178.9
9	1389.3	2571.8	1478.1	1388.4	1391.1
10	1329.6	2867.2	1407.4	1320.5	1341.7
11	1294.9	2149.8	1349.1	1277.8	1294.3

Tabla 3: Error Medio Cuadrático (factor de 10^6)

	Reg lin	Raz uni	Med uni	Reg rob	Corr rob
1	4.4737	378.264	4.6046	4.4916	4.4944
2	3.3879	59.064	3.5816	3.3884	3.4261
3	6.0839	95.802	7.1837	6.1490	6.0898
4	4.4690	27.124	4.5803	4.4728	4.4705
5	3.5925	88.007	3.6369	3.6223	3.6209
6	4.7441	124.846	496.74	4.7503	4.7813
7	3.5347	11.915	4.8793	4.0097	3.5371
8	5.1506	7.782	5.3685	5.1679	5.1825
9	7.0471	24.188	7.8753	7.0548	7.1119
10	6.3076	38.858	8.0901	6.4024	6.3345
11	7.0901	23.196	8.7311	7.2515	7.0901

La Fig. 5 muestra los gráficos de cuartiles (boxplots) correspondientes a los cinco métodos de normalización y el correspondiente a los conjuntos de datos originales (sin normalizar). En cada uno de los gráficos el eje de las abscisas muestra las distintas repeticiones del experimento, donde el primero (recuadro resaltado) representa el conjunto de referencia. El eje de las ordenadas representa el valor de las muestras.

Las rectángulos que aparecen en cada gráfico agrupan los datos de cada repetición entre el primer y el tercer cuartil, resaltando la mediana como punto intermedio. Específicamente, el lado inferior de cada pequeño rectángulo representa el primer cuartil, el lado superior el tercer cuartil de cada conjunto de datos, mientras que la línea intermedia dentro del rectángulo es el segundo cuartil, es decir, la mediana.

Cada gráfico de cuartiles muestra de forma general cuanta variación hay en los datos de las repeticiones normalizadas comparadas con el conjunto de datos de referencia, es decir, observando el gráfico de cuartiles de un método de normalización se puede estudiar la consistencia o no del proceso de normalización, ya que se espera que las distin-

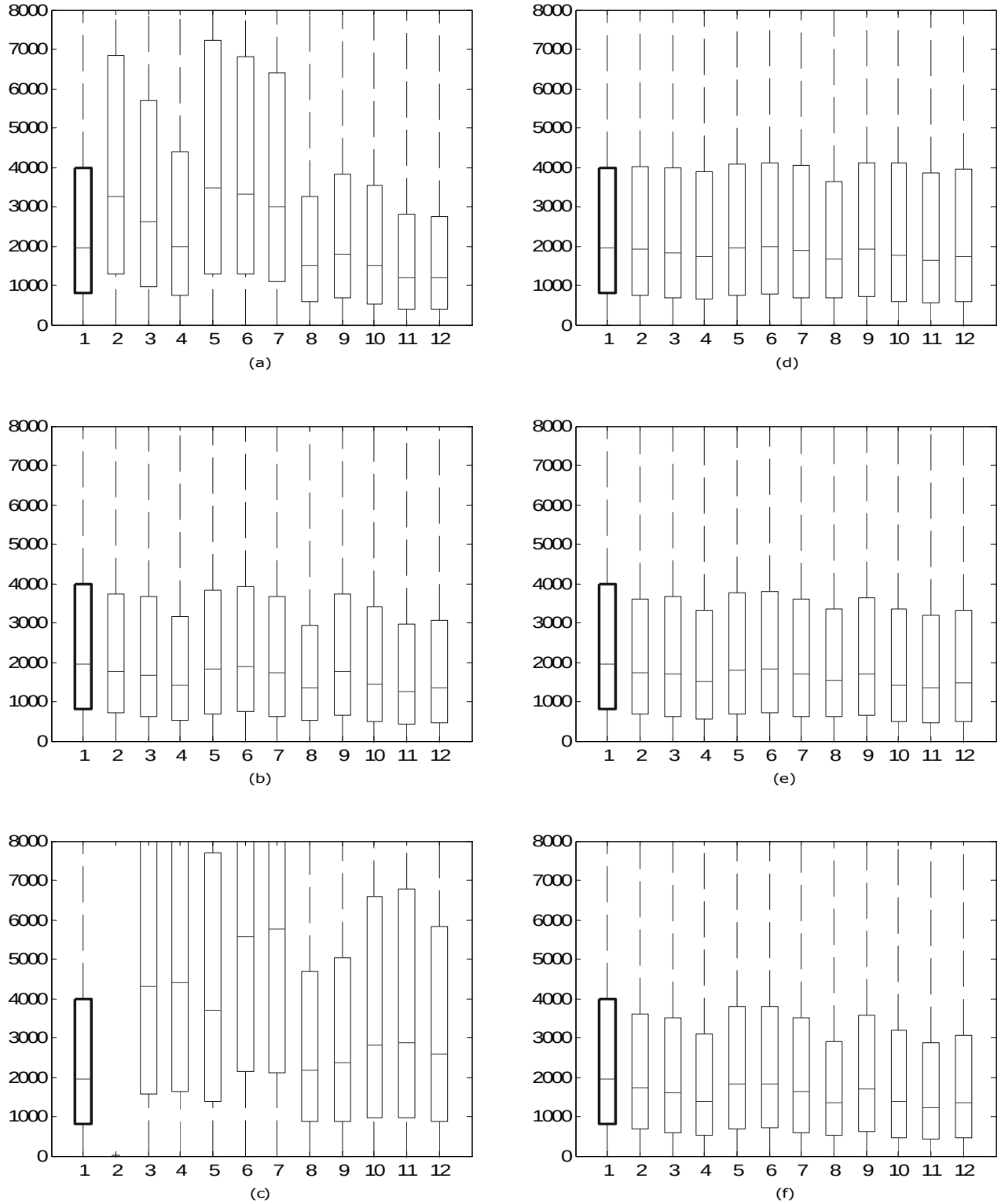

Fig.5: Gráfico de cuartiles de: (a) las repeticiones originales; (b) regresión lineal; (c) razón unitaria; (d) media unitaria; (e) regresión no lineal; y (f) correlación robusta.

tas repeticiones normalizadas den lugar a conjuntos de datos bastante similares.

Comparando los distintos gráfico de cuartiles se nota inmediatamente que las muestras sin normalizar son bastante dispersas, ya que la mediana de las distintas repeticiones varían en un rango dinámico muy amplio (alrededor de 2000), mientras que los conjuntos de muestras obtenidos a partir de la media unitaria y de la regresión robusta (figuras 5(d) y 5(e)) son los que menos variación muestran, tanto en la mediana como en el primer cuartil. Adicionalmente, como puede observarse en la Fig. 5(c) el método de normalización basado en la media de razón unitaria aumenta la dispersión de los datos y por tanto genera datos menos confiables. Por otro lado, los métodos de correlación robusta, Fig. 5(f), y de regresión lineal, Fig. 5(b), mejoran la dispersión de los datos originales, a pesar de no ser tan eficientes como los de media unitaria, y de regresión robusta.

7 Conclusiones

En el presente trabajo, se proponen dos nuevos métodos de normalización de datos de microarreglos de ADNc. Tales métodos estiman los parámetros de normalización bajo la suposición de que las repeticiones son proporcionales entre si y que los errores entre conjuntos de datos siguen una distribución de colas pesadas. Los métodos propuestos agregan robustez en la estimación del parámetro de normalización usando mediana ponderada como operador robusto. El primer método propuesto calcula el parámetro de normalización usando regresión lineal robusta, el cual tiene como objetivo la minimización del error medio absoluto. En cambio, el segundo método, propuesto como alternativa respecto al método de regresión lineal, estima el parámetro de normalización usando correlación robusta basada en mediana ponderada, en lugar de emplear correlación tradicional. Los métodos de normalización se evaluaron usando tres criterios de desempeño: el error medio cuadrático, el error medio absoluto y los gráficos de cuartiles. Desde el punto de vista del error medio cuadrático y del error medio absoluto, se observa que los métodos propuestos son competitivos respecto a los métodos lineales. Adicionalmente, los gráficos de cuartiles muestran que los datos normalizados usando los métodos propuestos, presentan una distribución similar a lo largo de las repeticiones. La robustez implícita en el operador de mediana evita que datos impulsivos originados en el proceso de adquisición del microarreglo produzcan estimaciones pobres de los parámetros de normalización.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Fondo Nacional de Ciencia, Tecnología e Innovación (FONACIT) bajo el proyecto Nro. 2005000234 y en parte por El Consejo de Desarrollo Científico, Humanístico y Tecnológico (CDCHT) de la Universidad de Los Andes bajo el proyecto Nro. I-768-04-02-B.

Referencias

Astola J y Kuosmanen P, 1997, Fundamentals of nonlinear digital filtering. CRC Press LLC, p. 276.

Arce G, 1998, A general weighted median filter structure admitting negative weights, IEEE Transactions on Signal Processing, Vol. SP-46, No. 12. pp. 3195-3205.

Arce G y Li Y, 2002, Median power correlation theory, IEEE Transactions on signal processing, Vol 50 N° 11, pp 2768-2776.

Bloch K y Arce G, 2002, Nonlinear correlation for the analysis of gene expression data, Proceedings of the 2002 Workshop on Genomic Signal Processing and Statistics, Raleigh, North Carolina.

Chen Y, Dougherty E y Bittner M, 1997, Ratio-based decisions and the quantitative analysis of cDNA microarray images, Journal of Biomedical Optics, Vol. 2, Nº 4, pp. 364-374.

Choi S, 2004, Dna chips and microarray analysis -an overview, Handbook of fungal biotechnology, Disponible en: http://www.its.caltech.edu/~schoi/DNA%20Chips% 20and %20Microarray%20Analysis%20-%20Sangdun% 20Choi.pdf.

Hedge P, Qi R, Abernathy K, Gay C, Dharap S, Gasparad R, Hughes J, Snesrud E, Lee N y Quackenbush J, 2000, A concise guide to cDNA microarray analysis, Biotechniques**,** vol. 29, pp. 548-557.

Lee MT, Kuo F, Whitmore GA y Sklar J, 2000, Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, Proceedings of the National Academy of Sciences, Vol. 97, Nº 18, pp. 9834-9839.

Mitra SK y Sicuranza GL, 2001, Nonlinear image processin, Academic Press, pp. 455.

Purdom E y Colmes S, 2005. Error distribution for gene expression data. statistical applications in genetics and molecular biology. Vol. 4, pp. 1-32. May. Disponible en: http://www.bepress.com/sagmb/vol4/iss1/art16

Rodriguez MA, 2005, Métodos robustos de estimación de correlación, Proyecto de grado, Facultad de Ingeniería. Universidad de los Andes, Mérida, Venezuela.