Análisis de microarrays, preproceso. Calidad en la selección de genes diferencialmente expresados

Mg. Nuria Ruiz Ruiz[1] anknur@gmail.com

Dr. Andrés Redchuk[1,2] andres.redchuk@urjc.es

Dr. Javier M. Moguerza[1] javier.moguerza@urjc.es

1. Departamento de Estadística e Investigación Operativa.
Universidad Rey Juan Carlos. Madrid, España.
2. Facultad de Ingeniería.
Universidad Nacional de Lomas de Zamora. Argentina.

Resumen

Como consecuencia del éxito de la tecnología de microarrays, aparecen en la literatura un gran número de experimentos realizados con los mismos. Sin embargo, los problemas de estandarización y las numerosas fuentes de variabilidad hacen necesarias técnicas de validación a posteriori. Por este motivo se ha tratado de estudiar cómo influye en la selección de genes diferencialmente expresados algunas de las principales técnicas de preproceso. Muchos de los estudios realizados para comparar estas técnicas, se han llevado a cabo sobre experimentos cuyos resultados óptimos se conocen a priori con el fin de intentar determinar cuál presenta mayor precisión. En nuestro caso no conocemos el resultado correcto a priori y lo que se ha realizado es un análisis comparativo de los resultados obtenidos en cada caso con el fin de poder predecir el comportamiento a priori de cada una de las técnicas analizadas en función de diversos factores como distribución de los datos iniciales, patrones de expresión objeto de interés, presencia de outliers, etc.

Se han aplicado tres técnicas de preproceso sobre un experimento de microarrays. Las técnicas aplicadas son GCRMA, MBEI y MAS5. Se han encontrado en nuestros datos principalmente tres patrones de expresión en aquellos genes diferencialmente expresados y se ha demostrado estadísticamente que existe una asociación entre la técnica de preproceso utilizada y el patrón predominante en la misma. Esta tendencia se ha relacionado con la eficiencia en la detección de valores atípicos y con la magnitud de cambio detectada con cada una de ellas. Por el momento, no se ha podido establecer un estadístico significativo a la hora de confirmar la concordancia entre los tres métodos tras la selección de genes diferencialmente expresados.

Palabras Clave: Microarrays, Optimización de procesos, Mejora de la Calidad.

Abstract

Following the success of microarray technology, in the literature there is a large number of experiments made with them. However, the problems of standardization and the many sources of variability make it necessary posteriori validation techniques. For this reason, we have tried to study how the selection of genes influence some key preprocessing techniques. Many of the studies conducted to compare these techniques have been carried out on experiments which optimal results are known a priori to try to determine which has greater accuracy. In our case, we do not know the correct result a priori and what has been accomplished is a comparative analysis of the results obtained in each case in order to predict a priori the behavior of each of the techniques discussed in terms of various factors as initial data distribution, expression patterns object of interest, the presence of outliers, and so on.

Three techniques have been applied on the preprocessing on a microarray experiment. The techniques are GCRMA, MAS5 and MBEI. In our data there have been found mainly three patterns of expression in those genes expressed and have been shown statistically that there is an association between the pre-processing technique used and the predominant pattern in it. This trend is related to efficiency in the detection of outliers and the magnitude of change detected with each of them. So far, it has not been able to establish a statistically significant in confirming the agreement between the three methods after the selection of differentially expressed genes.

Keywords: Microarrays, Process Optimization, Quality Improvement.

1. Introducción

Los microarrays de ADN pueden definirse como una matriz bidimensional de material genético que permite la automatización simultánea de miles de ensayos encaminados a conocer en profundidad la estructura y funcionamiento de la dotación genética, tanto en los diversos estados de desarrollo como en estados patológicos. Esta herramienta viene empleándose en dos aplicaciones fundamentales, la secuenciación o genotipado del ADN y el análisis de datos de expresión génica.

Los avances experimentados en técnicas de miniaturización y el desarrollo de la tecnología de imagen han provocado una rápida evolución de la tecnología de microarrays en las últimas décadas. Como consecuencia, si hace pocos años solamente podíamos estudiar la acción conjunta de unos pocos de genes, ahora podemos estudiar simultáneamente decenas de miles lo que hace necesario el empleo de herramientas estadísticas y matemáticas en el análisis de datos de microarrays. Se ha optado principalmente por el uso de técnicas de minería de datos ya existentes: técnicas de clustering y bioclustering, algoritmos genéticos, redes neuronales, herramientas de preprocesamiento de datos,

representación de modelos biológicos, biología computacional o minería de textos.

Sin embargo hasta la fecha, no se ha podido establecer un único estándar en esta tecnología sobre todo en lo referente a plataformas de fabricación, protocolos de ensayo y métodos de análisis de los datos obtenidos.

En lo que al análisis de datos se refiere aunque no existe un estándar podemos distinguir dos fases. La primera sería el preproceso que consta cuatro pasos. Los tres primeros (corrección del fondo, normalización y corrección de las PM sondas para los arrays de oligos) son opcionales. El último paso (sumarización) es imprescindible para obtener una medida del nivel de expresión génica. La segunda fase persigue como principal objetivo la obtención de genes diferencialmente expresados para posteriormente realizar diversos estudios en función de los objetivos del experimento u ensayo (clustering para determinar patrones de expresión, exploración funcional, etc.).

Algunas de las técnicas de preproceso mas utilizadas hoy en día son GCRMA (Robust Multichip Average with GC-content background Correction) (Wu et al., 2003), MBEI (A Model-Based Expression Indexes) (Li and Wong, 2001a) y (Li and Wong. 2001b) y MAS5 (Microarray Suite version 5.0) (Affymetrix, 2002) y (Sorin Draghici, 2003).

El objetivo de este trabajo es estudiar si influye y cómo influye cada una de estas técnicas de preprocesado en la selección de genes diferencialmente expresados así como estudiar la concordancia entre las mismas. Para ello se ha aplicado cada una de ellas al mismo conjunto de arrays y se ha procedido a la selección de genes diferencialmente expresados y al estudio de sus patrones de expresión utilizando el mismo procedimiento en cada uno de los casos. Posteriormente se ha llevado a cabo un análisis comparativo de los resultados obtenidos con cada uno de los tres métodos.

2. Material y método

2.1 Microarrays

Los chips empleados en este trabajo fueron Human Genome U133A 2.0 de la tecnología GeneChipTM de Affymetrix. Estos chips de oligonucleótidos permiten analizar el nivel de expresión de 18400 transcritos y sus variantes, incluyendo 14.500 genes humanos bien caracterizados, utilizando para ello 22.277 conjuntos de sondas y 500.000 oligonucleótidos distintos.

2.2 Diseño del experimento

Se ha utilizado un total de 20 chips, correspondientes a cuatro réplicas de cuatro tratamientos con ácido retinoico (RA), durante 30 minutos, 1, 6 y 24 horas, y a cuatro réplicas de control que no han recibido tratamiento alguno. Se realizaron añadiendo el compuesto al medio DMEM completo a una concentración final de 1 M. La línea celular de neuroblastoma humano es la SH-SY5Y crecidas en medio DMEM, suplementado con 10 % de suero fetal bovino, 2 mM de L-glutamina, 100 g/mL de penicilina y 100 g/mL de estreptomicina.

Esta línea se obtuvo de la European Collection of Cell Cultures (ECACC), nº 94030304.

2.3 Técnicas de preproceso

Como hemos dicho las técnicas de preproceso que nos interesa estudiar son GCRMA, MBEI y MAS5. El software utilizado en cada caso ha sido GeneSpring GX, DNA Chip Analyzer (dChip) y GenePat-tern respectivamente. A continuación se resumen los algoritmos utilizados por cada una de ellas en los distintos pasos de preprocesado.

GCRMA	gcrma	Quantiles		Median polish
MBEI		Invariant set	Subtract mm	LWR
MAS5	mas5	Scaling median	Ideal mm	mas5

2.4 Selección de genes diferencialmente expresados y detección de patrones de expresión

Dado que nuestro principal objetivo va a ser la selección de genes diferencialmente expresados para determinar posibles patrones de co-expresión vamos a utilizar la técnica de ANOVA (Análisis de la Varianza). En GCRMA y MBEI se ha aplicado el test de corrección para comparaciones múltiples False Discovery Rate (FDR) con el fin de reducir el número de falsos positivos considerando como genes diferencialmente expresados aquellos cuyo p-valor asociado es menor a 0,05. En MAS5 como sólo se habrían seleccionado 27 genes al aplicar la corrección FDR se ha optado por tomar 0,001 como punto de corte.

Para la determinación de patrones se han utilizado dos algoritmos de clustering aglomerativos, uno jerárquico (UPGMA) y otro no jerárquico (k-means). Se ha utilizado el coeficiente de correlación de Pearson para definir la distancia (ya que nos interesa estudiar los patrones de expresión) y se ha establecido previamente el numero de clases k igual a 3 en le caso de k-means tras varios ensayos. El software utilizado en este caso ha sido Gene Expression Profile Analysis Suite (GEPAS).

2.5 Análisis comparativo

Para analizar la asociación entre la técnica de preproceso empleada y el patrón de expresión predominante en el número de genes seleccionados como diferencialmente expresados se ha utilizado el estadístico Chi-cuadrado.

Para estudiar el nivel de asociación, se ha aplicado un análisis de correspondencias múltiples, específico para tablas de contingencias de orden m x n con m, n > 2. Se han seleccionado como variables de análisis el método de preproceso y el patrón de expresión. Los parámetros de dicho análisis son los siguientes:

- parámetro de ponderación = 1

- dimensión de la solución = 2
- método de normalización principal por variable
- estrategia para valores perdidos “Excluir los valores moda”.

Para estudiar el grado de concordancia entre los tres métodos se ha utilizado el coeficiente W de Kendall ya que N=3.
El software utilizado para este análisis ha sido SPSS y R.

3. Resultados

3.1 Calidad de los datos iniciales

La calidad del material genético utilizado se puede monitorizar mediante las denominadas gráficas de degradación. Las moléculas de ARN marcadas tienen dos extremos denominados 3' y 5'. Para construir estas gráficas de degradación, las sondas de cada conjunto son ordenadas en función de su posición relativa respecto al extremo 5' de la molécula de ARN marcada con fluorescencia. La degradación de la molécula de ARN normalmente comienza por el extremo 5', por lo que cabe esperar que la intensidad en este extremo sea inferior a la del extremo 3'. Mediante la representación de la media de intensidad calculada para cada posición sobre todos los conjuntos de sondas, se obtiene una idea de la degradación sufrida por el ARN. En nuestro caso y a la vista de las gráficas para nuestro experimento (fig. 1), el nivel de degradación de ARN parece seguir el comportamiento habitual. Se ha calculado el porcentaje de pares de sondas en los que la intensidad de la sonda MM es superior a la de la sonda PM. El resultado ha sido del 23.39%, que se encuentra ligeramente por debajo del porcentaje medio estimado del 30% (Naef and Magnasco, 2003).

Los datos de expresión pueden compararse usando las gráficas MAplot (Dalgaard, 2008). Este procedimiento es especialmente útil a la hora de diagnosticar problemas en cada conjunto de réplicas. En la fig. 2 se muestran todas las comparaciones entre pares de arrays llevadas a cabo en cada uno de los distintos conjuntos de réplicas. En el eje y se ha representado la diferencia de las intensidades de sonda entre los dos arrays en escala logarítmica (valor M), y en el eje x la media de las intensidades en escala logarítmica (valor A). Tal y como era de esperar, hay un buen ajuste entre réplicas de un mismo tiempo. Las diferencias más acusadas se dan entre las réplicas de 6 horas.

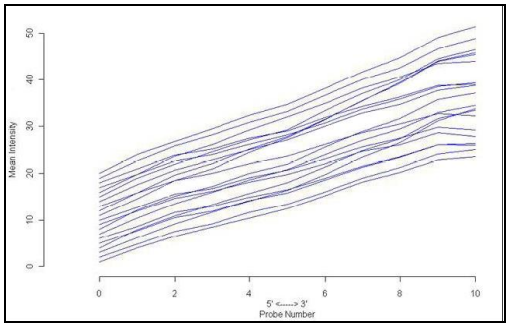


Fig. 1: Gráficas de degradación de ARN para cada array

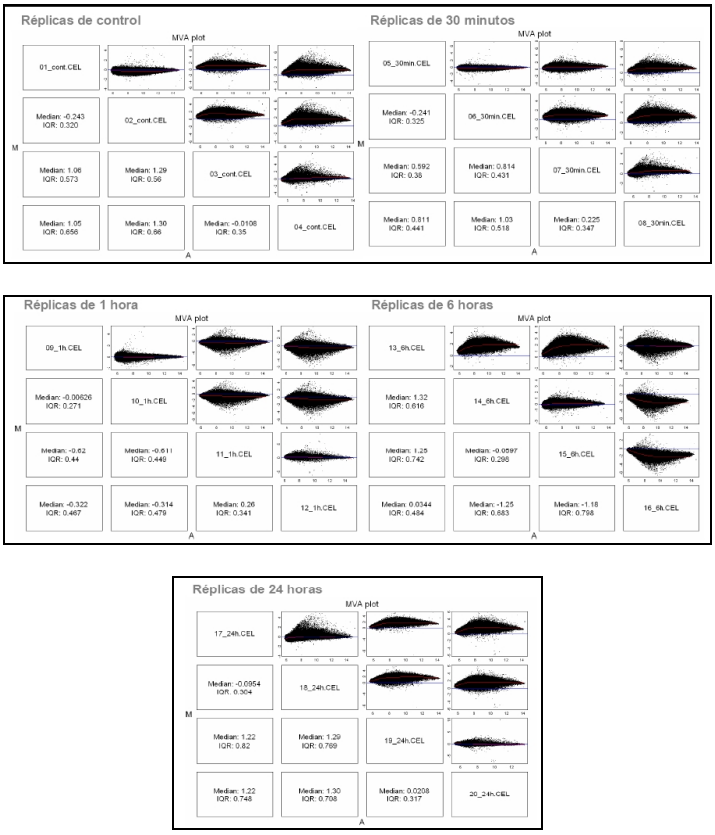


Fig. 2: MAplot de todas las comparaciones entre pares para cada grupo de réplicas

Por último, para tener una idea de la distribución de los valores de la intensidad en cada array, se muestra el diagrama de cajas en la fig. 3. Se pueden distinguir dos grupos. Aunque ambos muestran una distribución asimétrica, los datos de uno de ellos tienen mayor dispersión y los valores de intensidad son más elevados.

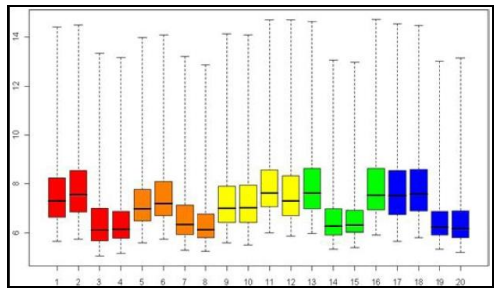
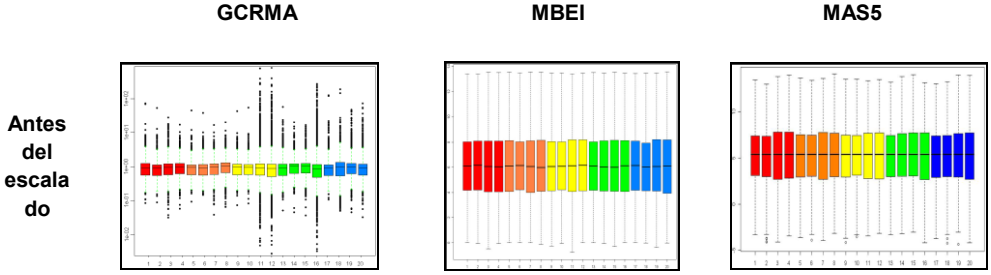


Fig. 3: Diagrama de cajas de los datos iniciales antes del preproceso. Cada caja se corresponde con un array. Se han utilizado los colores rojo, naranja, amarillo, verde y azul para agrupar las cuatro réplicas de cada tratamiento: control, 30 minutos, 1 hora, 6 horas y 24 horas, respectivamente.

3.2 Preprocesado

Tras aplicar cada una de las técnicas de preproceso se ha aplicado un escalado por mediana. Con este paso, el valor de expresión de los genes que apenas varía a lo largo de las distintas muestras se centrará a en una de ellas, lo que facilitará la detección de genes diferencialmente expresados. En la siguiente figura se muestran los diagramas de cajas antes y después del escalado para cada uno de los tres métodos.



Después del escalado

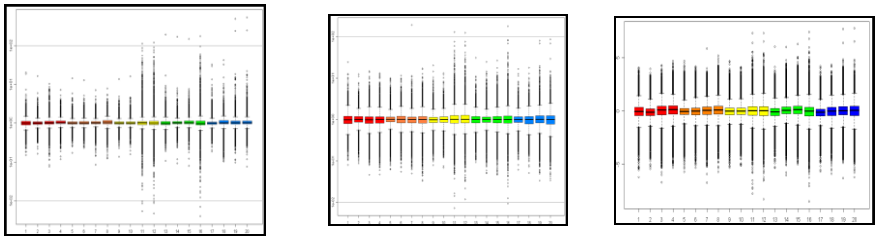


Fig. 4: Diagrama de cajas que muestra la nueva distribución de los valores de la intensidad en cada uno de los arrays tras el preproceso para cada una de las técnicas.

3.3 Selección de genes y de patrones de expresión

Como se puede apreciar en la fig. 5, los tres métodos han detectado principalmente tres patrones de comportamiento:

- Diferenciación temprana (genes que se han sobreexpresado a los 30 minutos). En la mayoría de estos genes el nivel de expresión vuelve a caer a las 24 horas.
- Diferenciación intermedia (genes que se sobreexpresan a las 6 horas de tratamiento)
- Diferenciación tardía (genes que se sobreexpresan a las 24 horas de tratamiento. La mayoría de estos genes se encuentran sobreexpresados en ausencia de tratamiento (muestras de control) por lo que podríamos decir también que su nivel de expresión decae a los 30 minutos de tratamiento.

GCRMA MBEI MAS5

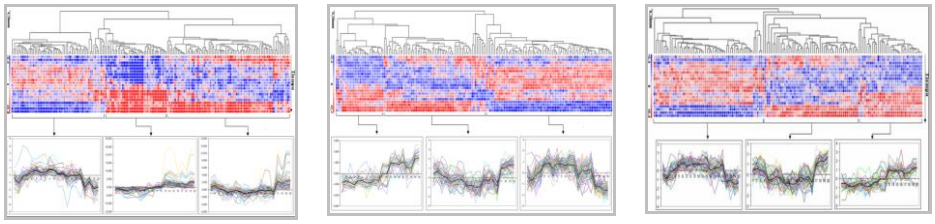


Fig. 5: Clusters obtenidos en cada técnica tras la selección de genes diferencialmente expresados. Los árboles se corresponden con los resultados obtenidos en UPGMA y las gráficas cartesianas con los resultados de k-means.

El número de genes de cada patrón para cada técnica se muestra en la siguiente tabla.

	Temprana	Tardía	Intermedia
GCRMA	56	81	56
MBEI	78	33	35
MAS5	47	59	47

Tabla 1: Número de genes seleccionado para cada patrón de expresión encontrado en cada una de las técnicas.

3.4 Análisis comparativo

Las tres técnicas han seleccionado un total de 339 conjuntos de sondas distintos, de los cuales 44 son comunes. De ellos, 14 corresponden a expresión temprana, 11 a diferenciación intermedia y 19 a expresión tardía. Hay también 65 conjuntos de sondas que han sido seleccionados en dos de los tres métodos. De ellos, hay cuatro genes cuyo tiempo de diferenciación varía de un método a otro. Estos cuatro genes se muestran en la siguiente tabla con el perfil de expresión asignado.

Conjunto de Sondas	Algoritmo	Tipo de diferenciación
201353_s_at	MBEI	Tardía
201353_s_at	MAS5	Intermedia
206638_at	GCRMA	Tardía
206638_at	MBEI	Intermedia
221796_at	GCRMA	Tardía
221796_at	MBEI	Intermedia
220610_s_at	GCRMA	Tardía
220610_s_at	MAS5	Intermedia

Tabla 2: Conjuntos de sondas con distinto patrón asignado por alguno de dos de los métodos.

En la tabla 3 se muestran los estadísticos descriptivos de cada array para los valores de la intensidad de los genes seleccionados. Si comparamos los estadísticos en cada grupo de tratamiento, puede observarse que en los grupos control, 30 minutos y 1 hora, los tres métodos presentan una media muy similar que se mantiene prácticamente constante durante los tres primeros tiempos de

tratamiento. En el método MAS5 se aprecia una mayor dispersión de los datos. A partir de las 6 horas de tratamiento, en los tres casos se produce un incremento de la media, la varianza y la desviación estándar siendo más acusado en el método GCRMA. Esto podría sugerir que los cambios más pronunciados en el nivel de expresión tienen lugar a las 6 horas de tratamiento, produciéndose los más leves por el mismo razonamiento en la fase temprana. En las dos últimas replicas de 24 horas también se produce un brusco aumento de la varianza y de la desviación estándar lo que puede sugerir la presencia de outliers en las mismas.

	Estadísticos		C1	C2	C3	C4	Valores medios	30min1	30min2	30min3	30min4	Valores medios	1h1	1h2	1h3	1h4	Valores medios	6h1	6h2	6h3	6h4	Valores medios	24h1	24h2	24h3	24h4	Valores medios	Valores medios totales
GCRMA	Media		1.001	1.017	0.966	0.959	0.986	0.936	0.961	0.947	0.954	0.950	0.952	0.951	1.024	1.013	0.985	2.503	2.262	2.133	1.728	2.157	1.701	1.910	6.122	6.524	4.064	2.68
	IC (95%)	Limite inferior	0.957	0.969	0.926	0.922	0.943	0.905	0.927	0.910	0.913	0.914	0.917	0.919	0.970	0.962	0.942	0.413	0.417	0.516	1.129	0.619	1.293	1.406	0.223	0.236	0.789	1.01
		Limite superior	1.046	1.065	1.008	0.997	1.029	0.967	0.995	0.989	0.995	0.986	0.988	0.983	1.078	1.065	1.028	4.594	4.108	3.750	2.328	3.695	2.109	2.415	12.022	12.811	7.339	3.15
	Mediana		0.986	0.952	0.977	0.998	0.978	0.955	0.960	0.982	0.984	0.979	0.931	0.953	1.002	0.973	0.965	1.022	1.010	1.072	1.015	1.030	1.174	1.344	1.357	1.463	1.334	1.25
	Varianza		0.092	0.108	0.081	0.065	0.087	0.044	0.054	0.066	0.078	0.061	0.059	0.047	0.136	0.122	0.091	203.158	158.328	121.527	16.685	124.924	7.746	11.835	1618.054	1837.739	868.843	205.06
	Desv. tip.		0.303	0.328	0.285	0.256	0.293	0.211	0.233	0.256	0.279	0.245	0.243	0.216	0.369	0.349	0.295	14.253	12.583	11.024	4.085	10.486	2.783	3.440	40.225	42.869	22.329	7.30
	Mínimo		0.314	0.328	0.201	0.140	0.246	0.318	0.246	0.148	0.123	0.209	0.317	0.305	0.282	0.160	0.266	0.332	0.442	0.527	0.295	0.399	0.463	0.354	0.368	0.391	0.394	0.36
	Máximo		2.370	0.440	2.842	2.291	1.986	1.540	1.598	1.713	2.076	1.732	1.779	1.756	3.249	2.919	2.426	192.439	170.063	149.125	47.707	139.834	32.135	41.538	489.583	525.269	272.131	90.92
	Amplitud intercuartil		0.340	0.417	0.223	0.206	0.296	0.219	0.285	0.185	0.262	0.238	0.298	0.274	0.345	0.352	0.317	0.489	0.436	0.281	0.537	0.436	0.762	0.849	1.289	1.378	1.069	0.54
	SEM		0.023	0.024	0.021	0.019	0.022	0.016	0.017	0.019	0.021	0.018	0.018	0.016	0.027	0.026	0.022	1.059	0.935	0.819	0.304	0.779	0.207	0.256	2.990	3.186	1.660	0.54
Dchip	Media		0.992	0.983	0.964	0.957	0.974	0.984	1.008	0.963	0.992	0.987	0.974	0.965	1.042	1.039	1.005	1.426	1.358	1.385	1.316	1.371	1.370	1.385	2.423	2.523	1.925	1.47
	IC (95%)	Limite inferior	0.944	0.941	0.924	0.920	0.932	0.952	0.972	0.928	0.954	0.952	0.939	0.932	0.989	0.990	0.963	1.063	1.069	1.072	1.101	1.076	1.158	1.161	1.230	1.271	1.205	1.22
		Limite superior	1.039	1.025	1.005	0.995	1.016	1.017	1.043	0.998	1.030	1.022	1.010	0.997	1.094	1.088	1.047	1.788	1.648	1.699	1.530	1.666	1.583	1.608	3.616	3.774	2.645	1.72
	Mediana		0.956	0.910	0.975	0.985	0.956	1.018	1.025	0.995	1.022	1.015	0.986	0.989	1.022	1.017	1.003	0.992	1.027	1.034	1.000	1.013	1.024	1.010	1.013	0.973	1.005	1.20
	Varianza		0.101	0.079	0.074	0.063	0.079	0.047	0.056	0.055	0.065	0.056	0.056	0.047	0.123	0.106	0.083	5.865	3.743	4.397	2.056	4.015	2.019	2.231	63.594	69.911	34.439	7.95
	Desv. tip.		0.318	0.281	0.271	0.250	0.280	0.218	0.236	0.234	0.255	0.236	0.236	0.217	0.350	0.325	0.282	2.422	1.935	2.097	1.434	1.972	1.421	1.494	7.975	8.361	4.813	1.65
	Mínimo		0.275	0.097	0.103	0.116	0.148	0.220	0.204	0.187	0.193	0.201	0.165	0.272	0.222	0.250	0.227	0.423	0.361	0.319	0.390	0.373	0.566	0.516	0.395	0.483	0.490	0.34
	Máximo		2.778	2.007	2.379	2.080	2.311	1.585	1.681	1.673	2.048	1.747	1.691	1.736	2.666	2.697	2.198	30.591	24.537	26.350	14.876	24.089	12.700	12.784	79.308	85.895	47.672	17.12
	Amplitud intercuartil		0.279	0.344	0.241	0.165	0.257	0.249	0.274	0.163	0.274	0.240	0.337	0.268	0.336	0.279	0.305	0.284	0.286	0.179	0.332	0.270	0.547	0.613	0.944	0.994	0.775	0.42
	SEM		0.024	0.021	0.021	0.019	0.021	0.016	0.018	0.018	0.019	0.018	0.018	0.016	0.027	0.025	0.021	0.184	0.147	0.159	0.109	0.149	0.108	0.113	0.605	0.634	0.365	0.13
MAS5	Media		0.930	0.865	1.257	1.263	1.079	1.006	0.952	1.135	1.199	1.073	0.942	0.959	1.177	1.226	1.076	1.322	1.474	1.932	1.172	1.475	0.977	1.130	2.592	2.775	1.869	1.55
	IC (95%)	Limite inferior	0.814	0.780	1.174	1.185	0.988	0.849	0.837	1.067	1.112	0.966	0.844	0.856	1.113	1.140	0.988	0.534	0.722	1.004	0.885	0.786	0.764	0.855	0.105	0.166	0.473	1.03
		Limite superior	1.045	0.949	1.341	1.263	1.169	1.163	1.067	1.203	1.286	1.180	1.039	1.062	1.240	1.313	1.164	2.110	2.225	2.860	1.460	2.164	1.191	1.406	5.079	5.383	3.265	2.07
	Mediana		0.743	0.747	1.241	1.340	1.017	0.882	0.795	1.159	1.159	0.999	0.806	0.826	1.102	1.185	0.980	0.762	1.024	1.186	0.992	0.991	0.784	0.814	1.177	1.292	1.017	1.20
	Varianza		0.558	0.300	0.291	1.244	0.598	1.030	0.555	0.193	0.313	0.523	0.397	0.445	0.171	0.311	0.331	25.945	23.593	36.020	3.452	22.252	1.900	3.167	258.507	284.438	137.003	33.33
	Desv. tip.		0.747	0.547	0.539	0.501	0.584	1.015	0.745	0.439	0.560	0.690	0.630	0.668	0.413	0.557	0.567	5.094	4.857	6.002	1.858	4.453	1.379	1.780	16.078	16.865	9.025	3.38
	Mínimo		0.188	0.131	0.080	0.157	0.139	0.140	0.093	0.141	0.087	0.115	0.264	0.219	0.241	0.104	0.207	0.103	0.099	0.178	0.079	0.115	0.102	0.126	0.095	0.239	0.141	0.17
	Máximo		5.479	4.730	4.059	4.330	4.649	10.293	6.902	4.773	4.426	6.599	5.284	6.187	3.052	4.604	4.782	64.893	62.729	75.999	24.270	56.973	16.648	20.266	206.173	216.247	114.834	41.22
	Amplitud intercuartil		0.370	0.232	0.416	0.399	0.354	0.250	0.315	0.278	0.369	0.303	0.322	0.286	0.293	0.326	0.307	0.370	0.196	0.298	0.236	0.275	0.219	0.454	0.360	0.460	0.373	0.38
	SEM		0.058	0.043	0.042	0.039	0.046	0.079	0.058	0.034	0.044	0.054	0.049	0.052	0.032	0.044	0.044	0.399	0.380	0.470	0.146	0.349	0.108	0.139	1.259	1.321	0.707	0.26

Tabla 3: Descriptivos para cada array en función de la intensidad de los genes seleccionados

Si observamos las frecuencias en la tabla 4 puede observarse las diferencias entre el porcentaje de genes seleccionados de cada patrón. La prueba Chi-cuadrado arroja un estadístico $p < 0.0001$ por lo que puede concluirse que existe una fuerte asociación entre la técnica de preproceso empleada y el patrón de expresión predominante en los genes seleccionados como diferencialmente expresados.

Tabla de contingencia Método x Patrón

			Patrón			Total
			Expresión temprana	Expresión intermedia	Expresión tardía	
Método	GCRMA	% dentro de Método	30,9 %	43,1 %	26,0 %	100 %
		% dentro de Patrón	29,0 %	53,4 %	30,7 %	36,8 %
	MBEI	% dentro de Método	46,8 %	19,1 %	34,1 %	100 %
		% dentro de Patrón	42,0 %	22,6 %	38,6 %	35,2 %
	MAS5	% dentro de Método	40,6 %	25,4 %	34,1 %	100 %
		% dentro de Patrón	29,0 %	24,0 %	30,7 %	28,0 %
Total		% dentro de Métodos	39,2 %	29,7 %	31,1 %	100 %
		% dentro de Curva	100,0 %	100,0 %	100,0 %	100,0 %

Tabla 4: Tabla de frecuencias. Se ha resaltado en amarillo oscuro los correspondientes niveles asociados a las variables método y patrón cuando coinciden con los porcentajes modales. Cuando no coinciden se encuentran resaltados en rojo.

Para estudiar el nivel de asociación, vamos a aplicar la técnica estadística de correspondencias múltiples al análisis de la tabla de contingencia. El primer resultado obtenido (tabla 5) es un análisis de homogeneidad que nos da una idea de la variabilidad explicada, aproximadamente del 56% en nuestro caso.

Resumen del modelo

Dimensión	Alfa de Cronbach	Varianza explicada		
		Total (autovalores)	Inercia	% de la varianza
1	0,375	1,232	0,616	61,575
2	0,042	1,021	0,511	51,072
Total		2,253	1,126	
Media	0,225	1,126	0,563	56,324

Tabla 5: Análisis de homogeneidad correspondiente al análisis de correspondencias múltiples

En el gráfico conjunto (fig. 6) se aprecia la tendencia de cada método al patrón correspondiente lo que coincide con lo observado en la tabla de contingencia. Los porcentajes de frecuencia para la variable método en el nivel GCRMA y el nivel intermedio de la variable patrón son moda para su columna y su fila respectivamente, lo que indica una fuerte tendencia en este caso. Lo mismo ocurre con los niveles MBEI y temprano de las variables método y patrón. Sin embargo en los niveles MAS5 y tardía de las variables método y patrón los porcentajes de frecuencia respectivos no son moda en su columna y en su fila correspondiente, lo que denotaría una asociación más débil (tabla 4).

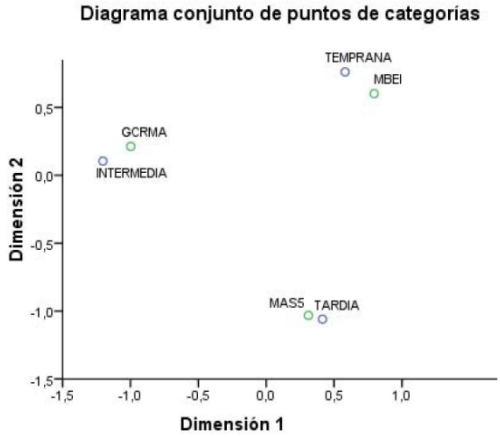


Fig. 6: Diagrama conjunto de puntos por categorías.

A la vista de los resultados obtenidos se puede afirmar que hay una relación directa entre el método utilizado en el preproceso y patrón dominante. Cabe preguntarse entonces cuál es el grado de concordancia entre los tres métodos. Es decir, queremos ver cuál es el grado de coincidencia de las “puntuaciones" asignadas por dichos métodos. Para ello se ha empleado el coeficiente de concordancia W de Kendall. Los resultados obtenidos se muestran en la tabla 6. El valor de la W obtenido (más cercano a cero que a uno) y la no significación de su estadístico asociado, no nos da evidencias estadísticas para rechazar la hipótesis nula de no concordancia.

Estadísticos de contraste

N	3
W de Kendall	0,312
Chi cuadrado	315,866
gl	338
Sig. asintót.	0,801

Tabla 7: Resultado del test de concordancia W de Kendall.

4. Conclusiones

Los análisis de expresión mediante microarrays ya son accesibles a la mayor parte de la comunidad científica. Los resultados obtenidos mediante estas técnicas están demostrando ser bastante reproducibles, además de aportar una gran cantidad de información sobre la regulación de la expresión génica en condiciones normales y patológicas. Como consecuencia del éxito de estas herramientas, aparecen en la literatura un gran número de experimentos realizados con arrays de expresión.

Sin embargo y hasta el momento, se hacen necesarias en este tipo de análisis técnicas de validación a posteriori. La causa principal de ello son los problemas de estandarización y las numerosas fuentes de variabilidad. De todas ellas, en lo relativo al tratamiento y análisis de datos, podemos mencionar las siguientes:

- La diversidad de procedimientos de procesamiento de datos.
- Sistemas inapropiados de análisis de datos.
- Escasez de recursos bioinformáticos para el tratamiento y comprensión de datos masivos.
- Por este motivo en este trabajo, se ha tratado de estudiar cómo influye en la selección de genes diferencialmente expresados algunas de las principales técnicas empleadas en el preproceso de los datos.

Hasta ahora, los esfuerzos encaminados a la comparación de métodos de preproceso se han llevado a cabo con el fin de determinar qué método presenta una mayor precisión. Para ello, se analizan experimentos de los que se conoce el resultado a priori. En nuestro caso el análisis comparativo se ha enfocado hacia la determinación de las diferencias o similitudes de los distintos subconjuntos de genes obtenidos para intentar determinar sus causas, con el fin de poder predecir el comportamiento a priori de cada uno de ellos en función de las características de nuestros datos iniciales y de los patrones de expresión que nos interese estudiar o que son objeto de nuestra búsqueda. Conocer este comportamiento nos resultaría útil a la hora de tomar una decisión sobre cuál de ellos utilizar. Aunque no sabemos a priori qué genes deberían haber sido

seleccionados como diferencialmente expresados, o cuál es el patrón de expresión que les corresponde en realidad, a la vista de los resultados obtenidos y en un intento de traducir los resultados a términos generales más allá de nuestro experimente podríamos formular las siguientes hipótesis:

En lo referente a la fase de normalización, el método que mejor parece uniformizar las distribuciones es GCRMA. Por el contrario, con el algoritmo MAS5 se consigue la menor uniformidad distribucional antes de la transformación a escala logarítmica.

El número de genes totales seleccionados como diferencialmente expresados es sensiblemente menor en MAS5 que en los otros dos métodos. Si se tiene en cuenta los estadísticos mostrados en la tabla 3 para las réplicas de 6 y 24 horas, y el aspecto comentado anteriormente acerca de la distribución obtenida tras la normalización, cabe pensar que este método sea menos eficiente en lo que a la detección de outliers se refiere.

En sentido opuesto al punto anteriormente mencionado, el método MBEI es el que mejor parece atenuar el efecto provocado por outliers, y el que mejor equipara la varianza de la intensidad entre réplicas. Este hecho nos hace pensar que sea el más restrictivo en lo que a eliminación de outliers se refiere. También parece que tiende a aumentar la varianza de las intensidades menores.

El método GCRMA parece amplificar bastante la varianza en los cambios más acusados del nivel de expresión.

Se puede observar en nuestro experimento que cada método parece detectar mejor los cambios significativos en un patrón distinto. Así, mientras que aplicando MBEI parecen detectarse mejor los cambios significativos en los genes de expresión temprana, con el GCRMA ocurre lo propio para genes de expresión tardía. Teniendo en cuenta los puntos anteriores, y en un intento de traducir este resultado a términos generales, se podría pensar que en el método GCRMA se tiende a seleccionar como patrón dominante el de aquellos genes cuyo cambio en el nivel de expresión presenta una varianza más acusada. Por otro lado, en el método MBEI se tiende a seleccionar como patrón dominante el de aquellos genes cuyo cambio en el nivel de expresión presenta una varianza menor.
No se ha podido establecer un estadístico significativo a la hora de confirmar la concordancia entre los tres métodos en la selección de genes diferencialmente

expresados, aunque dada la gran cantidad de genes que han sido descartados por no estar diferencialmente expresados (21938) se puede predecir una gran precisión a la hora de asignar una clasificación negativa.
Este último punto no significa que los tres métodos no sean concordantes, y nos sugiere una posible pauta a seguir en el intento de establecer los elementos que influyen en la clasificación de un gen como diferencialmente expresado y de la forma en la que lo hacen.
Dado que el algoritmo matemático empleado para la detección de genes diferencialmente expresados tras el preproceso de datos con los tres métodos que nos ocupan es el mismo, y que ANOVA resulta bastante útil a la hora de seleccionar genes en base a un estadístico y no de manera arbitraria, cabe la posibilidad de estudiar si los resultados obtenidos acerca de la concordancia y de la asociación se repiten aplicando otro tipo de método de clasificación que nos permita incluir de manera inequívoca un gen en un determinado grupo sin necesidad de un parámetro de entrada arbitrario, como ocurre en el caso del k-means (el método de agrupación más utilizado en la literatura). El procedimiento en cuestión podría ser un análisis factorial. Esto nos permitiría determinar de una forma más precisa el número de clusters más adecuado, y a la vez nos permitiría eliminar de los grupos de interés posibles patrones de expresión que se hayan introducido en los mismos de forma forzada debido al valor de k utilizado.

Si los resultados sobre concordancia y asociación no fueran los mismos con diferentes métodos de clustering, ello introduciría un nuevo factor más a tener en cuenta en la interpretación de los resultados. En cualquier caso y ya de forma general, los sucesivos avances en investigación en materia de microarrays aumentan la necesidad de que se produzcan avances tecnológicos en distintas vertientes especialmente relevantes para esta técnica. Sin embargo, los espectaculares avances experimentados en los últimos años en biotecnología, genómica, transcriptómica y, en general, en todas aquellas ciencias relacionadas con la genética, no van de la mano con los avances experimentados en ramas como la bioinformática, bioestadística, etc., que indudablemente se están convirtiendo en indispensables para la obtención de resultados y conclusiones. Las principales carencias a este respecto en las que se deberían enfocar los esfuerzos en este aspecto son las siguientes:

- Avance tecnológico en lo relativo a la potencia y optimización del hardware, sobre todo en lo referente a bases de datos. Aunque se pueden analizar simultáneamente cantidades ingentes de genes, la información obtenida necesita ser procesada y almacenada en bases de

datos, cuya estructura crece en complejidad a medida que lo hacen los avances biotecnológicos. Como consecuencia, se hace necesario disponer ordenadores y bases de datos cada vez más potentes. A este respecto, cabe destacar también la necesidad de unificación de las bases de datos ya existentes, así como de la estandarización de protocolos, nomenclaturas y resultados recogidos en las mismas.

- Revisión de los procedimientos existentes en el tratamiento de los datos, para descartar aquéllos que hayan quedado obsoletos.
- Creación de un protocolo estándar de preprocesado y análisis de datos, en el que se presenten los criterios más adecuados a seguir en función del objetivo de análisis, la estructura y distribución de los datos con el fin de optimizar los resultados.
- Por último, aunque imprescindible para los dos apartados anteriores, se requiere del desarrollo de nuevas técnicas matemáticas de validación que permitan calibrar la idoneidad del método utilizado, la evaluación de los errores cometidos y la comparación y toma de decisiones ante la variabilidad de resultados. En este sentido se están proponiendo nuevas posibilidades para la evaluación de distintas técnicas de bioclustering.

En resumen esta tecnología requiere de un mayor desarrollo en otras disciplinas y lo que es más importante, se necesita poder establecer validaciones adecuadas que permitan la traducción de todos los hallazgos realizados a la práctica, que es el principal objetivo en última instancia de dicha tecnología.

5. Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto Riesgos CM CAM s2009/esp-1594, y por el proyecto del Ministerio de Ciencia e Innovación, Hogar digital y contenidos Audiovisuales adaptados a los USuarios HAUS, IPT-2011-1049-430000.

Al Dr. José Manuel Romero Enrique, Profesor Titular de Universidad de Sevilla, por su tiempo y por todos los comentarios y consejos. Al Dr. Domingo Barettino Fraile, responsable de la Unidad de Biología de la Acción Hormonal del Instituto de Biomedicina de Valencia (CSIC) por ceder amablemente los datos de su experimento con microarrays y por toda la ayuda técnica prestada. A Manuel Soriano Cánovas, responsable informático del Instituto de Biomedicina de Valencia (CSIC) por facilitar todos los recursos informáticos necesarios y por su asesoramiento en esta área. A José Manuel Rojo Abuin, responsable de la Unidad de Análisis Estadístico del Instituto de Economía, Geografía y

Demografía del Centro de Ciencias Humanas y Sociales (CSIC) por el asesoramiento estadístico.

6. Referencias

[1] AFFYMETRIX, (2002): "STATISTICAL ALGORITHMS DESCRIPTION DOCUMENT". Technical Report, Affymetrix, Santa Clara, CA.

[2] ARIAS-CASTRO, E. CANDES, E. J., AND PLAN, Y., (2010): "GLOBAL TESTING AND SPARSE ALTERNATIVES: ANOVA, MULTIPLE COMPARISONS AND THE HIGHER CRITICISM". ar-Xiv:1007.1434.

[3] BREMER, M., HIMELBLAU, E. AND MADLUNG, A., (2010): "STATISTICAL METHODS IN MOLECULAR BIOLOGY". Methods in Molecular Biology, 620(3):287-313.

[4] BUTTE, A.J., (2009): "BIOINFORMATIC AND COMPUTATIONAL ANALYSIS FOR GENOMIC MEDICINE. ESSENTIALS OF GENOMIC AND PERSONALIZED MEDICINE, Chap. 10". Huntington F. Willard and Geo_rey S. Ginsburg, Elsevier, Academic Press.

[5] CALZA, S., VALENTINI, D., AND PAWITAN, Y., (2008): "NORMALIZATION OF OLIGONUCLEOTIDE ARRAYS BASED ON THE LEAST-VARIANT SET OF GENES". BMC Bioinformatics, 9(1):140.

[6] DALGAARD, P., (2008): "INTRODUCTORY STATISTICS WITH R". Springer, 2nd edition.

[7] DE BIN, R. AND RISSO, D., (2011). "A NOVEL APPROACH TO THE CLUSTERING OF MICROARRAY DATA VIA NONPARAMETRIC DENSITY ESTIMATION". BMC Bioinformatics, 12:49.

[8] DE HAAN, J.R., WEHRENS, R., BAUERSCHMIDT, S., PIEK, E., VAN SCHAIK, R. AND BUYDENS, L.M.C. (2007): "INTERPRETATION OF ANOVA MODELS FOR MICROARRAY DATA USING PCA". Bioinformatics 2007, 23:184-190.

[9] ELASHOFF, M., ALVARES, C. AND LAUREN, P., (2009): "METHOD AND SYSTEM FOR MANAGING AND QUERYING GENE EXPRESSION DATA ACCORDING TO QUALITY". US Patent 7,558,411, 2009.

[10] GAN, G., MA, C. AND WU, J. (2009): "DATA CLUSTERING: THEORY, ALGORITHMS, AND APPLICATIONS". SIAM, Society for Industrial and Applied Mathematics.

[11] IRIZARRY, R.A., WU, Z. AND JA_EE, H.A., (2006): "COMPARISON OF AFFYMETRIX GENECHIP EXPRESSION MEASURES". Bioinformatics 22, 789-794.

[12] LI, C. AND WONG, W. H., (2001A): "MODEL-BASED ANALYSIS OF OLIGONUCLEOTIDE ARRAYS: EXPRESSION INDEX COMPUTATION AND OUTLIER DETECTION". Proc Natl Acad Sci U S A, 98(1):31-36.

[13] LI, C. AND WONG, W. H., (2001B): "MODEL-BASED ANALYSIS OF OLIGONUCLEOTIDE ARRAYS: MODEL VALIDATION, DESIGN ISSUES AND STANDARD ERROR APPLICATION". Genome Biol, 2(8):RESEARCH0032.

[14] MUIR, W.M. ,ROSA, G.J.M., PITTENDRIGH, B.R., XU, Z., RIDER, S.D. , FOUNTAIN, M. AND OGAS J., (2009): “A MIXTURE MODEL APPROACH FOR THE ANALYSIS OF SMALL EXPLORATORY MICROARRAY EXPERIMENTS.COMPUTATIONAL”. Statistics and Data Analysis, 53:1566-1576.

[15] NAEF, F., AND MAGNASCO, M. O., (2003): “SOLVING THE RIDDLE OF THE BRIGHT MISMATCHES: LABELING AND ELECTIVE BINDING IN OLIGONUCLEOTIDE ARRAYS”. Physical Review E 68, 011906.

[16] NEPOMUCENO, J.A., TRONCOSO, A., AGUILAR-RUIZ, J.S., (2011): “BICLUSTERING OF GENE EXPRESSION DATA BY CORRELATION-BASED SCATTER SEARCH”. BioData Min., 4(1):3.

[17] RAUCH, G., GEISTANGER, A., TIMM, J., (2011): “A NEW OUTLIER IDENTI_CATION TEST FOR METHOD COMPARISON STUDIES BASED ON ROBUST REGRESSION”. Journal of Biopharmaceutical Statistics, 1520-5711, 21(1):151-169.

[18] Sorin Draghici, (2003): “DATA ANALYSIS TOOLS FOR DNA MICROARRAYS, 2A ED”. Chapman & Hall/CRC, 2003.

[19] WU, Z., IRIZARRY, R. A., GENTLEMAN, R., MURILLO, F. M., AND SPENCER, F., (2003): “A MODEL BASED BACKGROUND ADJUSTMENT FOR OLIGONUCLEOTIDE EXPRESSION ARRAYS”. Technical Report Working Paper1, Johns Hopkins University, Dept. of Biostatistics Working Papers.