**5 REFERENCIAS**

Ginebra, J. i Cabos, S. (1998). Anàlisi Estadística de l'estil Literari: Aproximació a l'autoria del Tirant lo Blanc. *Afers*, **29**, 185–206.

Riba, A. i Ginebra, J. (2000). Riquesa de Vocabulari i Homogeneitat d'estil en el Tirant lo Blanc. *Revista de Catalunya*, **13**, 99–118.

Riba, A. and Ginebra, J. (2005). Change-Point Estimation in a Multinomial Sequence and Homogeneity of Literary Style. *Journal of Applied Statistics*, (to appear).

Girón, J., Ginebra, J. and Riba, A. (2005). Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style. *The American Statistician*, **59**, **nº 1**, 1–12.

# ESTRUCTURA Y ANÁLISIS DE MICROARRAYS

**[1]Rivas-López, M.J., [1]Sánchez-Santos, J.M. y [2]De las Rivas, J.**
**[1] Dpto. Estadística. Universidad de Salamanca**
**[2] Centro de Investigación del Cáncer. Universidad de Salamanca**

A menudo hemos escuchado que los genes, o sus mutaciones, pueden ser muy influyentes a la hora de desarrollar una determinada enfermedad. El DNA existente en el núcleo de las células contiene las instrucciones para la fabricación de proteínas. Un gen es un segmento de DNA que contiene la secuencia codificante específica para construir una proteína concreta. Cuando un gen está activo en una célula, decimos que ese gen está "expresado" en ella. En una célula hay una gran cantidad de genes diferentes activos, que dan lugar a todas las proteínas que funcionan en ese tipo celular. Así por ejemplo, las células humanas epiteliales son diferentes a las células humanas musculares porque a nivel biomolecular los genes expresados y las proteínas presentes en ellas son distintos.

Como la actividad y expresión de los genes da información sobre los procesos biológicos y el comportamiento celular, tanto en estados normales como patológicos, su medición es importante para el avance científico. Hasta hace poco tiempo en los laboratorios de biología y genética molecular se podía medir la actividad de cada gen por separado, pero actualmente con las modernas técnicas de genómica puede medirse simultáneamente la actividad de decenas de miles de genes usando una herramienta nueva conocida como microarray de oligonucleótidos de DNA (Harrington et al. 2000). Los microarrays son dispositivos construidos usando microtécnicas de alta precisión capaces de sintetizar en superficies muy pequeñas (del orden de uno o varios $cm^2$) miles de copias de moléculas. En el caso de los microarrays para medir la expresión génica lo que se sintetizan son oligonucleótidos de DNA; es decir, fragmentos cortos de DNA. La información obtenida de los microarrays de expresión génica ayuda a contestar importantes preguntas biológicas y biomédicas como son:

- ¿Existen diferencias de actividad génica entre personas sanas y personas con una determinada enfermedad?
- ¿Existen subgrupos genéticos de personas con una determinada enfermedad que responden positivamente a un tratamiento específico?

## DISEÑO BASE DE UN MICROARRAY

El DNA es ácido desoxirribonucléico, una biomolécula formada por dos hebras ó cadenas cuyos eslabones son nucleótidos, cada uno incluyendo molecularmente un azúcar (la desoxirribosa), un fosfato y una base nitrogenada. Las bases nitrogenadas son de 4 tipos: adenina (A), citosina (C), guanina (G) y tiamina (T), de modo que cualquier DNA puede ser identificado específicamente por la secuencia lineal de las bases de sus nucleótidos, "secuencia genética", por ejemplo: ATTGCGCATA. De este modo, en

la secuencia reside la llamada “información genética” que es la que tienen los genes.
La tecnología utilizada en el diseño de los microarrays se apoya en la propiedad biomolecular fundamental del DNA que es la de "complementariedad” de las bases nitrogenadas, pues A y T, y C y G, se unen específicamente por puentes de hidrógeno. Es decir, si una porción de un DNA presenta en una hebra la secuencia TGAAACT se puede deducir que la hebra complementaria de ese fragmento de DNA será necesariamente la secuencia ACTTTGA.
Resumiendo lo anterior, la doble cadena de DNA está constituida por dos hebras que tienen secuencias de nucleótidos complementarias y se unen por las bases nitrogenadas de cada hebra dando lugar a una estructura de doble-hélice, especie de escalera helicoidal, que es tan característica del DNA (Figura 1). En cada escalón o peldaño de la estructura del DNA aparecen ligadas dos bases complementarias. Así pues, si desligamos la escalera de DNA nos encontramos con dos hebras simples, cada una complementaria de la otra.

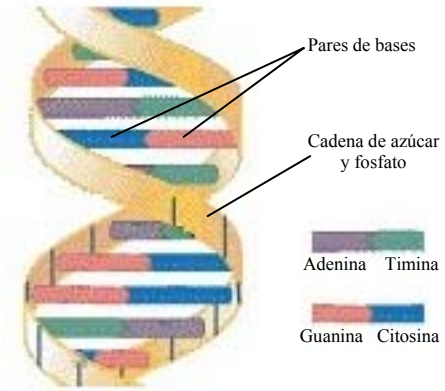


Figura 1: Doble-hélice de DNA

Todas las células de nuestro cuerpo tienen en su núcleo el mismo contenido genético, constituido por el conjunto de moléculas de DNA que es propio de cada especie. En nuestro caso, especie humana *Homo sapiens*, son 23 moléculas distintas de DNA duplicadas, es decir 46 moléculas. Cada molécula cuando se condensa y compacta se puede ver al microscopio como una unidad llamada cromosoma. Todo el conjunto de moléculas de DNA es lo que se llama genoma. Sin embargo, cada tipo de célula tiene activado un tipo de genes del genoma y esa activación corresponde a los llamados “genes expresados”, es decir genes que han sido transcritos por la maquinaria de trascripción celular a mRNA (moléculas de ácido ribonucleico mensajero), que son copias de una hebra del DNA génico. Los mRNA salen del núcleo y se van a traducir, por la maquinaria celular de traducción, en proteínas. Por tanto, podemos determinar qué genes están expresados o activados en una célula midiendo la cantidad de mRNA correspondiente a ese gen que hay en ella. Sin embargo, el mRNA libre es muy inestable, por lo que para poder manipularlo la biotecnología ha diseñado herramientas para pasarlo a DNA y esos DNA que provienen de “retro-trascripción” in vitro de mRNA se llaman cDNA.
En 1996 se comercializaron los primeros microarrays de DNA que, aunque con gran variedad de formas, tienen el mismo diseño base. Un microarray es un "chip" del tamaño aproximado de un sello donde, sobre una matriz inerte, se depositan miles de “hebras simples” de DNA de secuencias génicas, oligonucleótidos, de modo que sobre ellos pueden hibridar las secuencias complementarias correspondientes que son las que se obtienen del mRNA de las muestras biológicas que queremos analizar.

Aunque hay diferentes técnicas para la construcción de microarrays de DNA (Schena et al. 1995; Lockhart et al. 1996), el procedimiento básico para trabajar con ellos es el siguiente:

1. Marcar la muestra del tejido a estudiar con un tinte fluorescente.
2. Aislar el mRNA de las células de interés y proceder a copiarlo mediante una síntesis *in vitro* para pasarlo a cDNA.
3. Desnaturalizar ese cDNA para obtener hebras simples.
4. Poner el cDNA troceado sobre el microarray donde las hebras simples de cDNA son atraídas por las hebras simples de oligonucleótidos del microarray uniéndose a ellas para volver a conformar la estructura de doble-hélice similar a la del DNA (proceso conocido como *hibridación*).

5. Lavar el microarray para quitar las hebras simples de la muestra que no han hibridado.
6. Escanear el microarray con un láser para cuantificar la fluorescencia de cada gen.

En general, la actividad de un gen está representada por el número de copias de mRNA de ese gen en una muestra de células. Un alto (bajo) nivel de fluorescencia indica que muchas (pocas) copias del mRNA de ese gen han hibridado en el chip y que, por tanto, el gen tiene mucha (poca) actividad en la célula.

A. Malcolm Campbell del Davidson College, ha realizado una animación del proceso del microarray de DNA que puede encontrarse en:

*http://www.bio.davidson.edu/courses/genomics/chip/chip.html*

## ANÁLISIS DE DATOS DE MICRO-ARRAYS CON MAS5.0 (AFFYMETRIX)

Nos centraremos en un tipo específico de microarrays como son los microarrays de oligonucleótidos, de los cuales los más famosos son los microarrays GeneChip de la compañía americana *Affymetrix*. Un oligonucleótido es una cadena de nucleótidos, que en el caso de los microarrays de *Affymetrix* son 25. Un microarray de este tipo consiste en una matriz inerte sobre un substrato de vidrio en donde se han sintetizado miles de cadenas de oligonucleótidos de secuencias distintas. A su vez, el microarray está dividido en celdas donde se colocan miles de copias de cada tipo de oligonucleótido.

Para cada gen, el microarray contiene un conjunto de sondas o "probe-set" que son oligonucleótidos de 11 tipos distintos (así es en el caso del microarray modelo U133 de genes humanos); esto es, 11 hebras simples distintas de oligonucleotidos de tamaño 25. Cada oligonucleótido está miles de veces pegado en una celda constituyendo una sonda o "probe". Ahora bien, para cuantificar de alguna manera el número de "falsas" hibridaciones de estas hebras simples cada celda está dividida en dos partes: una con copias del oligonucleótido correcto y otra al lado con ese oligonucleótido modificado en la base central de su secuencia. Estos dos tipos de oligos constituyen un "probe-pair" y el que tiene la cadena correcta, se llama "perfect match" (PM), y el que tiene la cadena alterada se llama "mismatch" (MM) (Figura 2).

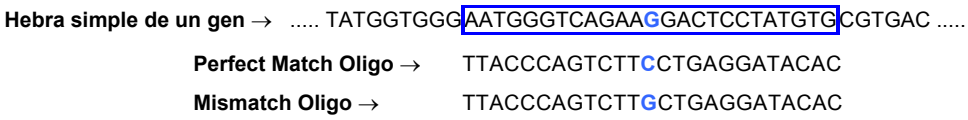


Figura 2: Probe-pair para un gen

Antes de comenzar el análisis de los datos hay que depurarlos pues hay muchas fuentes de variación: diferentes lecturas de los niveles de fluorescencia (dependiendo del escáner utilizado), microarrays con cantidad variable de tinción fluorescente, etc. Este proceso de depuración se lleva a cabo mediante el ajuste, escalado y normalización de la señal de fluorescencia.

El *ajuste de la señal* de fluorescencia se realiza restando a la fluorescencia inicial una estimación de la fluorescencia debida al fondo (background) del microarray. El *escalado* consiste en multiplicar las intensidades del microarray "muestra" por un factor de escalado (SF) para que su media de intensidad sea igual a una intensidad predeterminada. La *normalización* se utiliza cuando se pretende comparar varios microarrays "muestra" entre sí, y consiste en multiplicar las intensidades de cada microarray "muestra" por un factor de normalización (NF) para que sus medias de intensidad sean iguales a la de un microarray "referencia". Si sólo tenemos una "muestra" y la queremos comparar con la "referencia" el NF será 1. Para los procesos de escalado y normalización se considera como media de intensidad de un microarray a la media recortada excluyendo el 2% menor y mayor de las intensidades.

## ANÁLISIS DE DATOS DE UN ÚNICO MICROARRAY

Si nos proponemos estudiar un gen en un único microarray, el MAS5.0 (Microarray Analysis Suite 5.0) de Affymetrix nos informa sobre:

1. Detección del gen (gen presente/ausente)
2. Señal de intensidad (medida de la expresión del gen)

### 1. Detección del gen

Para cada "probe pair" j (1, 2, …, 11) del "probe set" que representa al gen, se calcula el cociente $Rj = (PMj - MMj) / (PMj + MMj)$, y se compara con un umbral predeterminado ($\tau$) para evitar falsas presencias (que se producen cuando *MMj* y *PMj* son parecidos); de modo que a las diferencias $(Rj - \tau)$ se les aplica el test unilateral de los rangos con signo de Wilcoxon, generándose así un *p-valor de detección del gen*. Es decir, este test nos indica al final si un "probe set" concreto correspondiente a un gen se puede tomar como "presente" (P) o "ausente" (A) en la muestra analizada.

### 2. Señal de intensidad

Se pretende cuantificar la fluorescencia de cada "probe-pair" y dar un valor de expresión del "probe-set"; es decir, la señal de intensidad del gen. Los pasos en la construcción de la señal de un gen i son:

1.- La intensidad total de cada "probe-pair" *j* del gen *i* es la diferencia de la dada por la hibridación "verdadera" del gen y la "dispersa" ó debida a otras causas. Aunque debería poder estimarse por *PM – MM*, no es así pues pueden aparecer intensidades negativas cuando *MM > PM*. Por ello se define el Ideal Mismatch (*IM*) de cada "probe-pair" como el *MM si MM < PM* y como una modificación del *MM si MM ≥ PM*, de modo que las diferencias *PM - IM* sean siempre positivas.

2.- Se calcula el valor de cada "probe-pair" *j* del gen *i*, denominado "probe value" (*PV*), siendo: $PV_{ij} = log_2 (V_{ij})$ donde $V_{ij} = max \{ PM_{ij} - IM_{ij} , 2^{-20} \}$. El logaritmo de la señal correspondiente al gen i (*Signal Log Value (i)*) será la media biponderada de Tukey de sus 11 "probe values": $Signal\ Log\ Value\ (i) = Tbi\ (PV_{i1}\ ,\ \dots\ ,\ PV_{i11})$.

3.- La señal final correspondiente al gen *i* es: $Signal\ (i) = NF \cdot SF \cdot 2^{SignalLogValue(i)}$.

Es decir, nos da un valor absoluto de intensidad de señal para cada "probe set" concreto, correspondiente a un gen, en números que oscilan normalmente entre 1 y 10000.

## ANÁLISIS DE DATOS DE DOS MICROARRAYS, UNO DE MUESTRA Y OTRO DE REFERENCIA

Si ahora queremos estudiar el cambio experimentado en la señal de un gen en dos microarrays "muestra" y "referencia", el MAS5.0 de Affymetrix nos informa sobre:

1. Detección de cambio en la señal (genes que cambian y en qué sentido)
2. Señal relativa (medida de la expresión relativa de cambio)

### 1. Detección de cambio en la señal

Para cada gen se calculan tres vectores, $v[k]$, $k = 0,1,2$, de longitud 22 a partir de los valores de los 11 "probe-pairs" que lo representan en los microarrays "muestra" y "referencia" y de un rango para los factores de normalización, por defecto de ±10%.

Para cada $k = 0, 1, 2$, se realiza el test unilateral de los rangos con signo de Wilcoxon y se extiende a un contraste bilateral, con lo que se obtienen tres p-valores, *p0*, *p1* y *p2*, que se utilizan para definir un *p-valor de cambio del gen*, *p*, bastante restrictivo a la hora de considerar que existe cambio (incremento ó disminución) en la expresión del gen en la "muestra" respecto a la "referencia".

Es decir, este test nos indica para cada "probe set" concreto si su señal disminuye respecto al control ($p \sim 1$), si su señal se incrementa respecto al control ($p \sim 0$), ó si no hay cambio significativo en la señal ($p \sim 0.5$). También se admite la distinción de disminución marginal (*MD*) o de incremento marginal (*MI*) cuando el p-valor esté muy cerca del que delimita la región de aceptación

de disminución o incremento respectivamente.

**2. Señal relativa**

Se pretende cuantificar el cambio en la señal de intensidad del gen. Los pasos en la construcción de esta señal relativa para un gen i son:

a.- Para cada "probe-pair" *j* del gen *i* se calcula su señal escalada y normalizada $SPV_{ij} = log_2 (SF \cdot NF \cdot V_{ij})$ donde $V_{ij} = max \{ PM_{ij} - IM_{ij} , 2^{-20} \}$, y como señal relativa del "probe-pair" se toma la diferencia *PLRij = SPVij (muestra) – SPVij (control)*.

b.- La señal relativa correspondiente al gen *i* (*Signal Log Ratio (i)*) será la media biponderada de Tukey de las señales relativas correspondientes a sus 11 "probe-pairs": (*Signal Log Ratio (i)*) = $Tbi (PV_{i1} , \ldots , PV_{i11})$.

Es decir, nos da para cada "probe set" concreto (correspondiente a un gen) un valor relativo de cambio de intensidad de señal en escala 2-logarítmica, en números que oscilan normalmente entre -4 y 4.

## PROBLEMAS A LA HORA DE REALIZAR EL ANÁLISIS ESTADÍSTICO DE MICROARRAYS

Un problema a la hora de analizar datos procedentes de microarrays es que el número de genes es mucho más grande que el número de individuos. Es decir en términos estadísticos tenemos el caso de miles de variables frente a sólo unas pocas muestras. Este tipo de datos hace imposible invertir las matrices de dispersión y por tanto los métodos de reducción de variables como la regresión han de desestimarse pues usan dichas inversas para calcular las estimaciones mínimo-cuadráticas. Aunque se han desarrollado algunas técnicas de reducción de datos, hay aún que desarrollar métodos para averiguar qué conjunto de genes es el más informativo.

Como a menudo estamos interesados en estudiar la variación significativa de ciertos genes particulares, usamos técnicas de contraste gen-a-gen como t-tests, ANOVA ó regresión. Estas técnicas presentan el problema grave de que cada vez que un gen es considerado significativo en alguna de estas pruebas, se produce un *error de Tipo I*. Si predeterminamos un nivel de significación de 0.05 y ejecutamos un t-test individual sobre 10000 genes, entonces cabe esperar que 500 genes sean considerados significativos, aunque ni tan siquiera tengan señal alguna. Aunque el control de este tipo de errores se había estudiado ampliamente (Benjamini y Hochberg, 1995), se ha retomado en el contexto específico de datos de microarrays (Storey, 2002).

Otro problema es que los datos de microarrays no verifican las suposiciones usuales de muchos tests estadísticos estándar. Suelen presentar asimetría por la derecha y varianzas desiguales. Considerando los datos log-transformados conseguimos mejorar la asimetría pero las varianzas siguen siendo bastante desiguales, lo cual hace que muchas técnicas (como el ANOVA) no sean robustas. Por ello se está investigando sobre transformaciones y normalizaciones de datos de microarrays que permitan que los análisis estadísticos estándares sean fiables (Durbin et al, 2002).

## ¿QUÉ SOFTWARE EXISTE PARA EL ANÁLISIS DE MICROARRAYS?

El software para realizar análisis de datos de microarrays está siendo desarrollado constantemente. Algunos de estos programas y métodos disponibles gratuitamente son:

• Bioconductor: Conjunto de programas y aplicaciones gratuito que trabaja con el lenguaje R y que precisa de conocimiento básico de programación en R ó S-Plus. Está diseñado específicamente para extraer información de microarrays y tiene muchos tipos de gráficos (*www.bioconductor.org*).

• SAM & PAM: Programas gratuitos para Microsoft Excel ó R. El SAM descubre genes significativos, controlando la tasa de falsos positivos, y el PAM los clasifica mediante métodos centróides (*www-stat.stanford.edu/~tibs*).

• BRB ArrayTools: Este software está diseñado para Microsoft Excel y se utiliza para la visualización y el análisis estadístico de datos de microarrays. Realiza

comparación y predicción de clases y tests de permutaciones para los niveles de significación (*http://linus.nci.nih.gov/BRB-ArrayTools.html*).

**REFERENCIAS**


Affymetrix 2002. Statistical algorithms description document (MAS v5.0).

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society,* B, 57:289-300.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. 2002. A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics,* 18:105S-110S.

Hardin, J. 2005. Microarray data from a statistician´s point of view, *STATS,* 42:4-13.

Harrington, C.A., Rosenow, C. and Retief, J. 2000. Monitoring gene expression using DNA microarrays, *Current opinion in Microbiology*, 3:285-291.

Lockhart, D., Dong, H., Bryne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology,* 14:1675-1680.

Schena, M., Shalon, D., Davis, R., and Brown, P. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science,* 270:467-470.

Storey, J. 2002. A direct approach to false discovery rates, *Journal of the Royal Statistical Society,* B, 64:479-498.


# 2. ARTÍCULOS DE INVESTIGACIÓN OPERATIVA

## MÁQUINAS DE VECTOR DE APOYO: PROBLEMAS DE PROGRAMACIÓN MATEMÁTICA EN CLASIFICACIÓN


**Emilio Carrizosa y Belén Martín-Barragán**
**Dpto. Estadística e Investigación Operativa. Universidad de Sevilla**
**ecarrizosa@us.es,belmart@us.es**


### 1. Introducción

En la última década, la capacidad de almacenamiento de información digital se ha duplicado cada nueve meses. Crece, por tanto, a una velocidad muy superior a la prevista por la ley de Moore para el crecimiento de la capacidad de cálculo, [18, 25], provocando la aparición de las denominadas *fosas de datos*, [18]: datos que son almacenados y descansan en paz, sin que nadie los reclame o los recuerde.

La constatación de la existencia de tales fosas de datos, y la consiguiente pérdida de oportunidades de avance en el conocimiento o de negocio, está provocando un enorme interés por el desarrollo de técnicas que, complementando a las previamente existentes, permitan obtener información desconocida y potencialmente útil de datos provinientes de campos tan diversos como la Bioinformática (expresión genética,…), gestión de clientes (fuga de clientes, análisis de la cesta de la compra,…), la banca (valoración de riesgo en créditos, detección de uso fraudulento de tarjetas de crédito, …), Internet (clasificación de páginas web, filtrado de correo indeseado, …), [1, 2, 3, 16, 19, 20, 22, 35].

Hablamos, usando una denominación de moda en los medios científicos, y, en particular, en las líneas editoriales de algunas de las revistas de más alto índice de impacto en nuestra área de conocimiento, de la