

00. Página de la materia

http://www.dm.uba.ar/materias/optativas/aspectos_estadisticos_de_microarreglos/2007/2

Objetivos del curso

- Desarrollar en el/a alumno/a la capacidad de analizar en forma crítica los métodos de análisis de datos de experimentos de microarreglos propuestos.
- Realizar comparaciones de entre los métodos utilizados en los inicios de estas tecnologías, los actuales y las nuevas propuestas, teniendo presente sus supuestos y limitaciones.
- Discutir publicaciones seleccionadas.
- Realizar prácticas con las principales propuestas utilizando el entorno de R y paquetes de Bioconductor.

Programa

1. Revisión de temas de biología molecular. Nivel de expresión de genes. Oligos. Sensibilidad y especificidad para un gen.
2. Microarreglos de uno y dos canales.
3. Diseño del experimento. Fuentes de sesgo. Aleatorización. Controles locales.
4. Diseño de las sondas (probes) y diseño de las muestras dentro de cada arreglo. Controles potenciales para normalización.
5. Diseño entre arreglos. Tipos de muestras. Replicación, técnica, biológica. Muestras individuales vs. muestras combinadas. Muestras combinadas vs. muestras amplificadas.
6. Revisión de procedimientos básicos en R. Estructura de datos en R. Estructura de datos de microarreglos. Bioconductor.
7. Representación gráfica de un experimento. Comparaciones directas, indirectas y diseño de loop. Criterios de evaluación del diseño: Robustez extensibilidad y eficiencia. Eficiencia teórica vs. resultados empíricos.
8. Medidas del nivel de expresión de los genes. Cuantificación y normalización en datos microchips de un canal. Corrección por intensidad del fondo.
9. Comparación de métodos de selección de genes candidatos a estar expresados diferencialmente.
10. Clusters para datos de microarreglos. Diagnóstico y validación.
11. Clasificación.
12. Análisis de genes corregulados: Gene set enrichment analysis.

Bibliografía

Analyzing Microarray Gene Expression Data. G. McLachlan, K. Do, C. Ambroise. Wiley 2004.

Bioinformatics and Computational Biology Solutions Using R and Bioconductor Editado por R. Gentleman, V. Carey, W. Huber, R. Irizarry, y S. Dudoit (2005). Springer.

Statistical Analysis of Gene Expression Data. Editado por T. Speed. (2003). Chapman&Hall

Microarray Analysis. M. Schena (2003). Wiley

0. Introducción

Los experimentos con microarreglos forman parte de lo que se ha dado en llamar Bioinformática. En ésta se integran métodos matemáticos, estadísticos y de ciencias de la computación, para analizar datos que provienen de la biología molecular.

Los experimentos requieren de las siguientes etapas:

- diseño
- mediciones
- preprocesamiento y
- postprocesamiento

En esta primera clase veremos:

- algunos temas de biología molecular que son indispensables para comprender los experimentos de microarreglos y su posterior análisis de datos
- los principales tipos de microarreglos.
- cómo se realiza un experimento típico de microarreglos
- consideraciones sobre diseño específicas para este tipo de tecnologías.

1. Algunos temas de biología molecular

1.1 Ácidos nucleicos (DNA y RNA)

La estructura básica de los ácidos nucleicos son los *nucleótidos*. Podemos clasificar a los ácidos nucleicos en dos tipos:

DNA formado por una *doble* cadena de *nucleótidos*
RNA formado por una *simple* cadena de *nucleótidos*

1.1.1 Nucleótidos

En la figura 1

- El azúcar es una molécula, esquematizada por un pentágono, de 5 carbonos cuyas posiciones se indican con 1', 2', 3', 4', 5'.
- La base nitrogenada está unida al carbono en la posición 1' del azúcar mediante una unión covalente
- El fosfato (ácido fosfórico se encuentra enlazado por una unión éster fosfato) está unido con la posición 5' del azúcar.
- En 3' el nucleótido tiene un OH (oxhidrilo) libre.

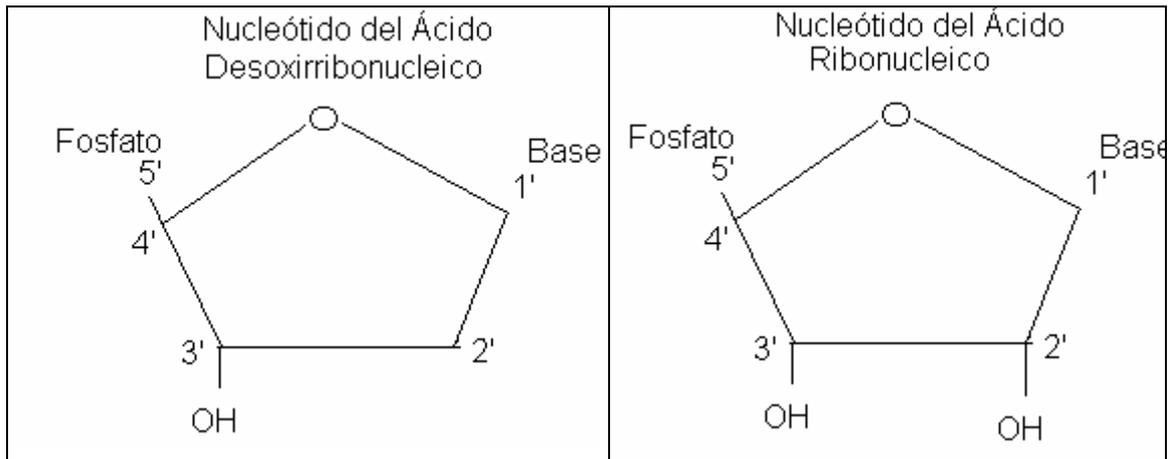


Figura 1. Estructura química de un nucleótido

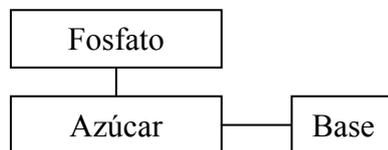


Figura 2. Esquema un nucleótido

Cada nucleótido está compuesto por

- fosfato
- azúcar (en RNA es *ribosa*, en DNA *desoxirribosa*)
- base nitrogenada que puede ser una de las siguientes:

Adenina (A)	en DNA y RNA
Citosina (C)	en DNA y RNA
Guanina (G)	en DNA y RNA
Timina (T)	en DNA
Uracilo (U)	en RNA

1.1.2 Cadena de polinucleótidos

Los nucleótidos se unen formando una cadena de polinucleótidos (figura 3). Tanto en los DNA como en los RNA la unión se realiza mediante un enlace entre el grupo 5' fosfato de un nucleótido y el grupo 3' oxhidrilo del azúcar de otro nucleótido. Un extremo del polímero de ácido nucleico tiene un oxhidrilo libre (el **extremo 3'**), el otro extremo posee un fosfato (el **extremo 5'**).

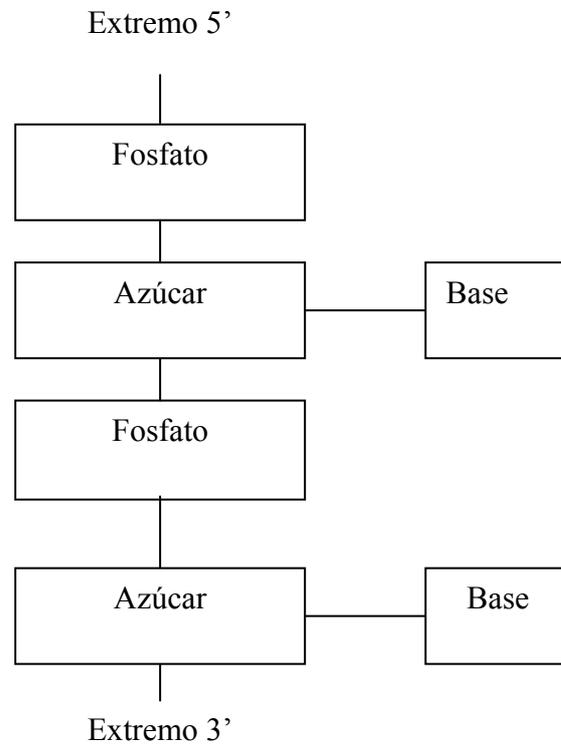


Figura 3. Cadena simple de nucleótidos

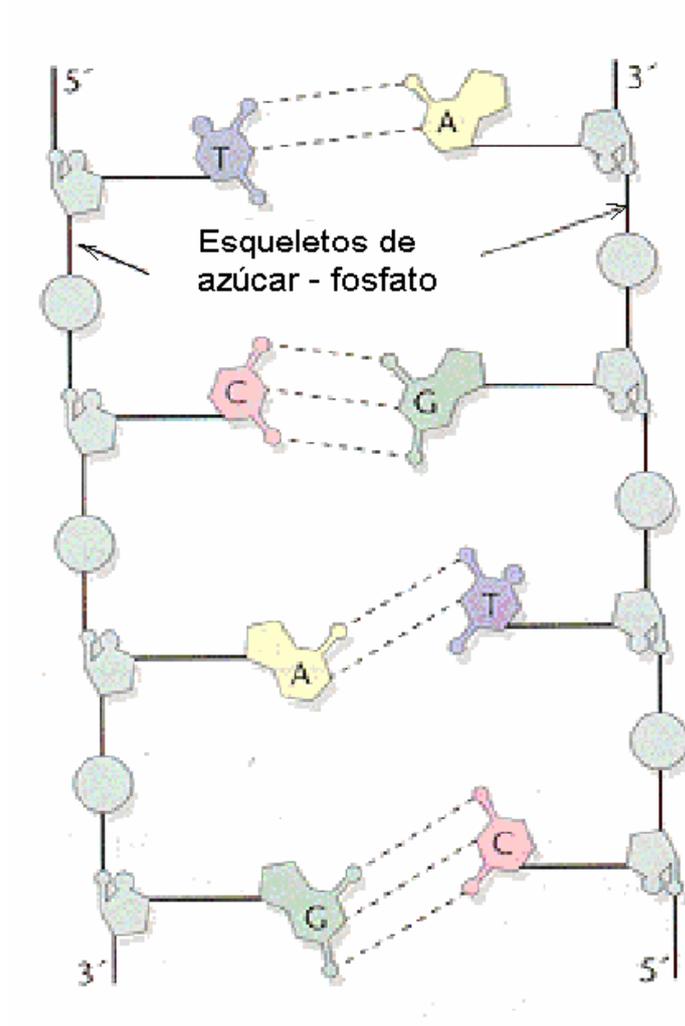
1.1.3 *Oligonucleótidos*

Los oligonucleótidos son secuencias cortas de nucleótidos de RNA o DNA. Estas secuencias pueden tener unos 20 o menos bases o pares de bases. Muchas veces los oligonucleótidos son referidos simplemente como *oligos*.

Cuando la secuencias son de 50-70 nucleótidos hablamos de oligonucleótidos *largos*.

1.2 *Estructuras*

1.2.1 *Estructura del DNA*



La estructura del DNA consiste de una *doble cadena* de polinucleótidos unida por (puentes hidrógenos entre) las bases de acuerdo con la siguiente regla complementaria

$$\begin{aligned} C &\equiv G && \text{con 3 puentes de hidrógeno} \\ A &= T && \text{con 2 puentes de hidrógeno} \end{aligned}$$

de acuerdo con el modelo de James Watson y Francis Crick (1953).

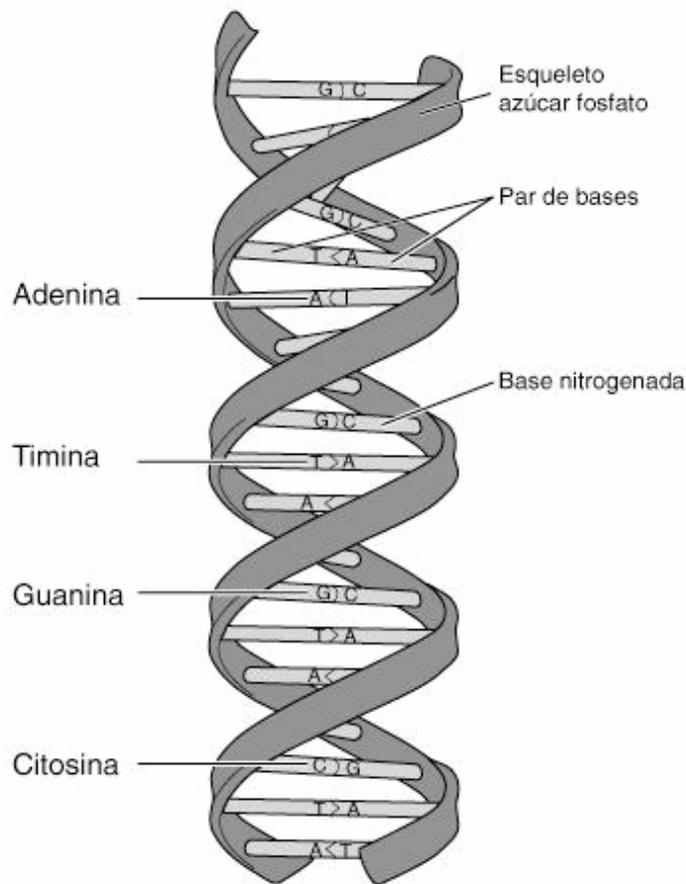


Figura 4. Doble cadena DNA

En 1962 James Watson (1928–), Francis Crick (1916–2004) y Maurice Wilkins (1916–2004) recibieron en forma conjunta el Premio Nobel de Medicina por su determinación en 1953 de la estructura del ácido desoxirribonucleico. Rosalind Franklin (1920–1958), quien murió de cáncer a los 37 años no pudo recibirlo.

Este modelo además postula que la molécula de DNA consiste de dos hebras de polinucleótidos enroscadas una alrededor de la otra en forma de *doble hélice*, como una escalera helicoidal con el esqueleto de azúcar-fosfato del lado de afuera y las bases hacia dentro. De manera que una base de una hebra apunta hacia la base de la otra hebra. Volviendo a la analogía de la escalera, el esqueleto de azúcar-fosfato vendría a formar los costados de la escalera y entre medio las bases (unidas por enlaces de hidrógeno) los peldaños. Cada hebra del DNA es la mitad de la doble hélice. Las dos mitades se juntan en una estructura de doble hélice (figura 4).

1.2.2 Estructura de RNA

El RNA es un polinucleótido de *simple cadena* con las mismas bases que las del DNA salvo que la Timina (T) es reemplazada por el Uracilo (U) y como ya hemos visto el azúcar es *ribosa* y no *desoxirribosa* como en el DNA

Para más detalles: Química Orgánica. John McMurry

1.3 La célula

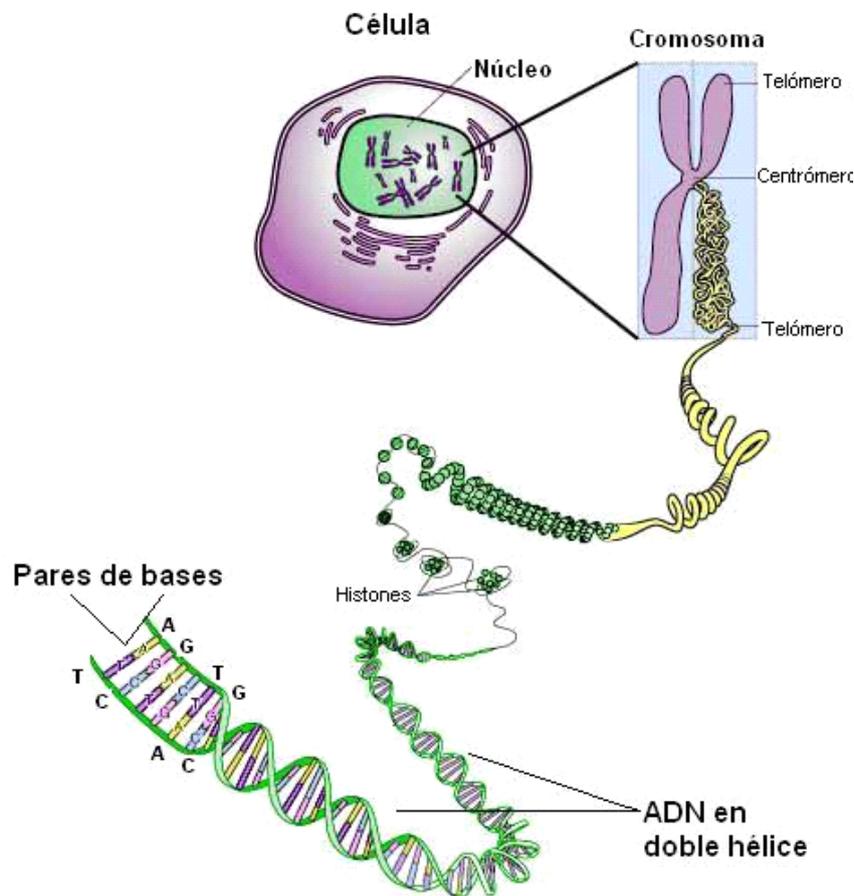


Figura 5. Una célula eucariota y un cromosoma aumentado

La figura 5 muestra el esquema de una célula eucariota, es decir con un núcleo y un citoplasma separados por una membrana (las células procariotas no tienen núcleo diferenciado, por ej. en bacterias). Dentro del núcleo se distinguen los cromosomas. Un cromosoma se encuentra aumentado para destacar su estructura de DNA. La molécula de DNA consiste en una doble cadena complementaria. Como ya hemos mencionado cada cadena consiste de una estructura de fosfatos y azúcares que sostienen diferentes

secuencias de cuatro posibles bases (Adenina Guanina Citosina y Timina). Cada una de ellas puede unirse mediante un puente de hidrógeno a su base complementaria según la regla descubierta por Watson-Krick (A-T C-G) como ya hemos visto.

1.4 Genoma Humano

El genoma de una célula es su contenido total de DNA. En las células eucariotas el DNA se encuentra en las mitocondrias y en el núcleo (DNA nuclear). Nos ocuparemos solamente de este último.

El núcleo de toda célula humana contiene 46 cromosomas (23 pares). En cada cromosoma se encuentra una larga cadena de DNA formada por aprox. 3×10^9 pares de bases, esta larga cadena desenroscada puede medir hasta 12 cm.

En la siguiente dirección es posible hallar las secuencias de DNA de diversos organismos

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=genome>

Todas nuestras células contienen *la misma información genética*. ¿Qué es lo que hace que por ejemplo las células de la piel sean diferentes de las del hígado?

Estas diferencias resultan del hecho que diferentes genes se expresan en diferentes niveles.

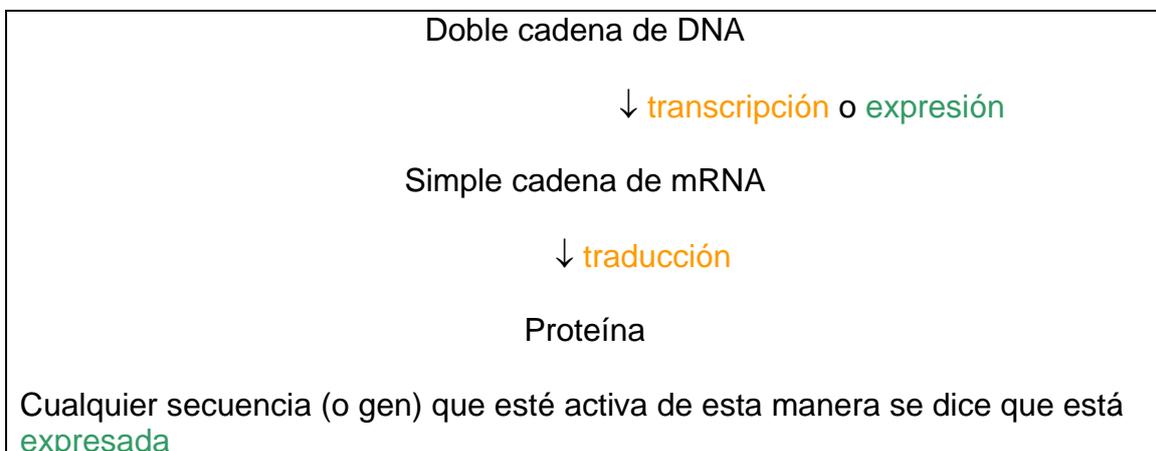
Aquí tenemos dos nuevas preguntas:

- 1) ¿Qué es un gen?
- 2) ¿Qué significa que un gen se exprese?

Llamaremos gen a un segmento específico de la molécula de DNA que contiene toda la información necesaria para instruir a la célula que sintetice un producto específico.

Para la segunda pregunta pasamos a la siguiente sección

1.5. Dogma central de la biología molecular



El dogma de la biología establece que una porción del DNA del cromosoma se copia (transcripción) a una cadena simple de mRNA (RNA mensajero) que sale del núcleo llevando consigo la información necesaria para codificar (traducción) una proteína.

Dentro de cada porción del DNA, que llamaremos gen, hay segmentos que conocemos tienen un papel activo en el proceso de codificación (exones, es la parte del mRNA que sale del núcleo luego de la transcripción) y también hay otros segmentos que no codifican (intrones parte del mRNA que se transcribió pero que no sale del núcleo).

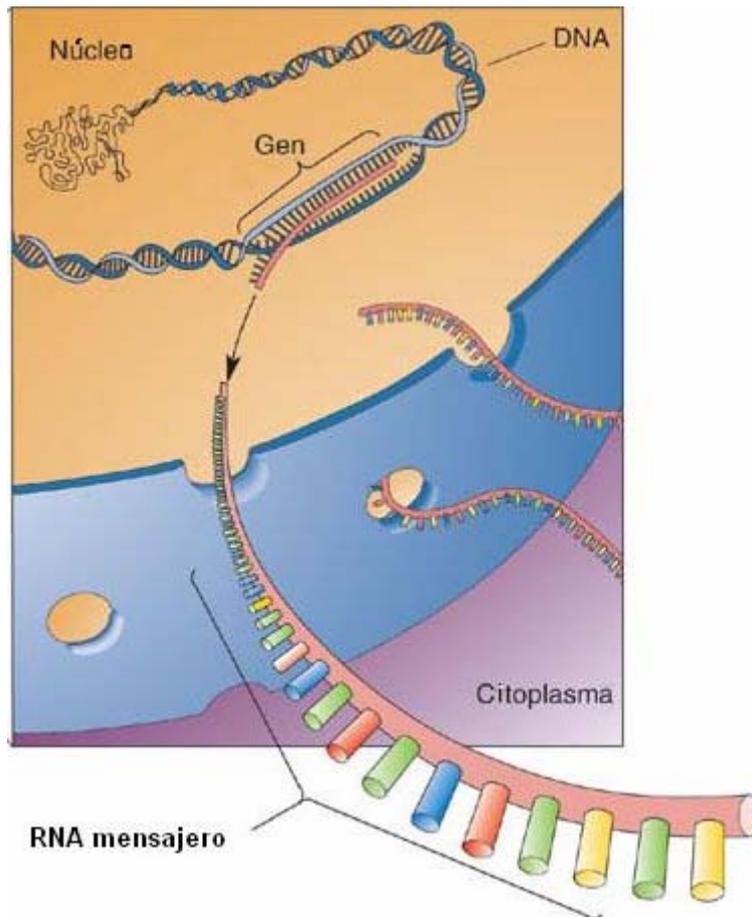


Figura 6a. Transcripción de un gen

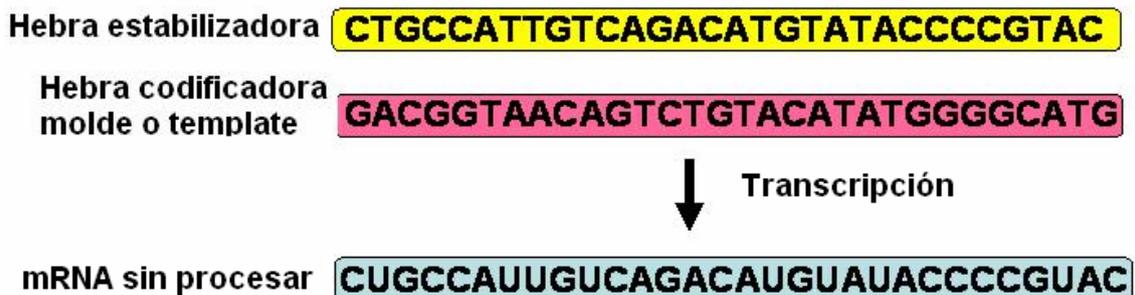


Figura 6b. Transcripción de un gen

En el proceso de transcripción, se sintetiza en forma complementaria una secuencia de RNA mensajero a partir de una secuencia de bases de la hebra de DNA que actúa como molde. De esta manera la secuencia de bases de RNA es igual a la hebra de DNA estabilizadora pero cambiando la T por la U.

El RNA mensajero que sale del núcleo sólo tiene los exones y en general es más corto que la porción de DNA que lo codificó y por lo tanto más corto que la hebra estabilizadora. Es el mRNA *maduro* que ha sufrido los procesos de capping (G), polyadenylation (AAAA...) y splicing.

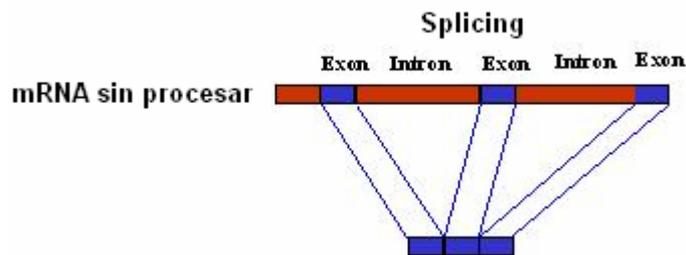


Figura 6c. Esquema del proceso de splicing

Ojo! En el esquema de la figura 6a no se pone de manifiesto que el la secuencia que se ha transcripto es más larga que la secuencia de mRNA maduro que sale del núcleo.

Cualquier secuencia (cadena genómica, o gen) que esté activa de esta manera se dice que está *expresada*, como se muestra en las figuras 6a, 6b y 6c.

El *nivel* de expresión de una un gen es la *cantidad de copias* de mRNA transcritos presentes en la célula en un determinado momento

Perfil de expresión

Si pudiésemos contar la cantidad de moléculas de mRNA para cada gen en una única célula obtendríamos su perfil de expresión “verdadero”. La figura 7 muestra un perfil de expresión “verdadero” hipotético.

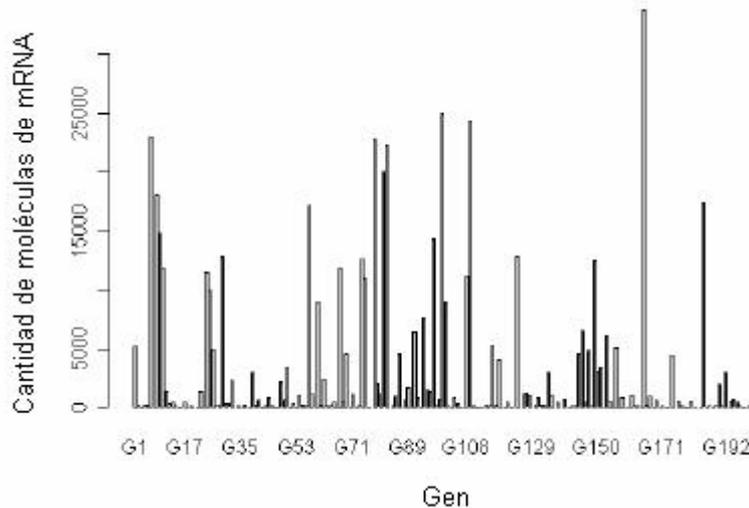
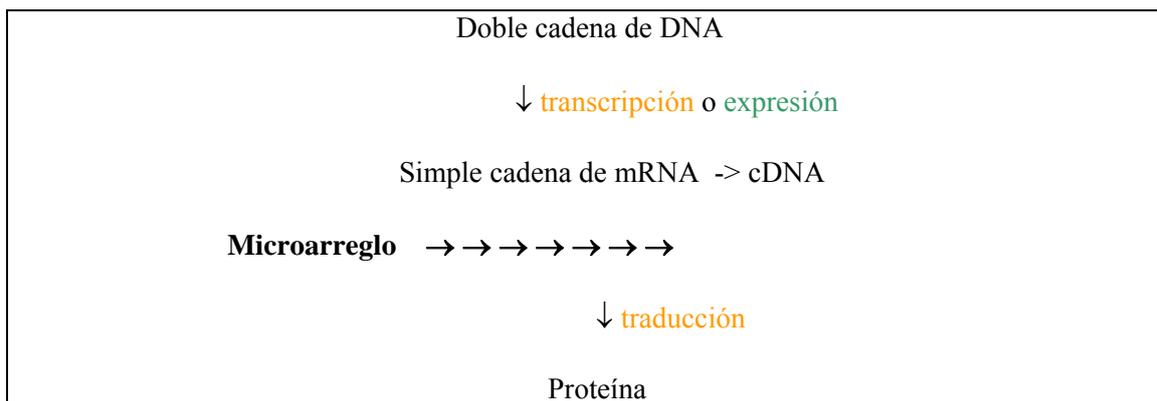


Figura 7. Perfil de expresión “verdadero”

No describiremos el proceso de síntesis de las proteínas que dirige el RNA mensajero, pero sí debemos tener en cuenta que la cantidad de proteína sintetizada es proporcional a la cantidad de RNA mensajero transcrito. Es esa cantidad de un RNA mensajero transcrito, presente en la célula en un determinado momento, que llamamos **nivel de expresión de un gen**.

2 Microarreglos

En un experimento de microarreglos se interrumpe el proceso natural determinado por el dogma de la biología se extrae el mRNA maduro de uno o más tejidos para hibridarlo (veremos más adelante qué significa esto) con el material que se encuentra previamente depositado sobre el microarreglo. El microarreglo actúa como un detector de la cantidad de RNA mensajero presente en el tejido.



La figura 8 muestra el gráfico correspondiente a la cantidad de publicaciones que contienen la palabra clave “Microarrays” por año obtenidas utilizando el buscador de PUBMED.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed>

Desde que Schena M, et al. Publicaron Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (1995) el crecimiento de la

cantidad de publicaciones por año fue exponencial hasta el 2001 y a partir de allí es lineal con un incremento de aproximadamente 1000 publicaciones más cada año.

Cantidad de publicaciones por año

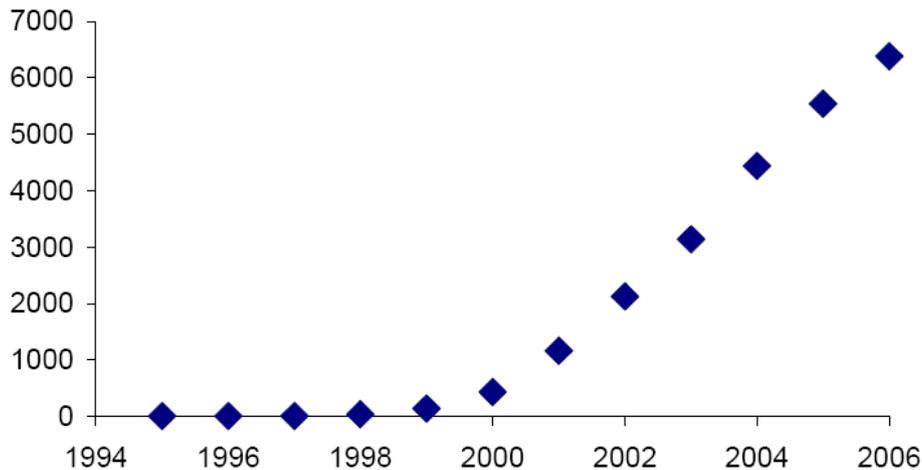


Figura 8

La tecnología de los Microarreglos ha abierto la posibilidad de medir el nivel de expresión de miles de secuencias genómicas simultáneamente en una gran variedad de organismos y a cualquier momento de su desarrollo. Tales experimentos producen a su vez gran cantidad de datos genéticos que pueden ser utilizados para realizar preguntas biológicas o médicas. Esto ha generado muchas expectativas en el avance de los conocimientos sobre:

- procesos moleculares biológicos
- diagnóstico y pronóstico de enfermedades (pronóstico=diagnóstico precoz)
- mecanismos acción de una droga
- mejoramiento de las estrategias terapéuticas

2.1 ¿Qué es un microarreglo?

Es un soporte sólido, generalmente vidrio o silicio, al que se le han adherido, mediante un robot, en forma ordenada sondas (probes) con diferentes cadenas conocidas de material genético (DNA, cDNA, oligos) (cubriendo parte o toda la secuencia de un genoma-transcriptoma de un organismo), en forma matriz de miles de puntos (10000 – 40000) equiespaciados. Cada secuencia se asocia con un único gen (tiene alta especificidad para ese gen). Cada punto contiene millones de secuencias clonadas “idénticas”.

- **Cada punto contiene millones de clones de una secuencia específica, asociada a un gen.**
- **Se puede o no tener el conocimiento sobre la secuencia.**

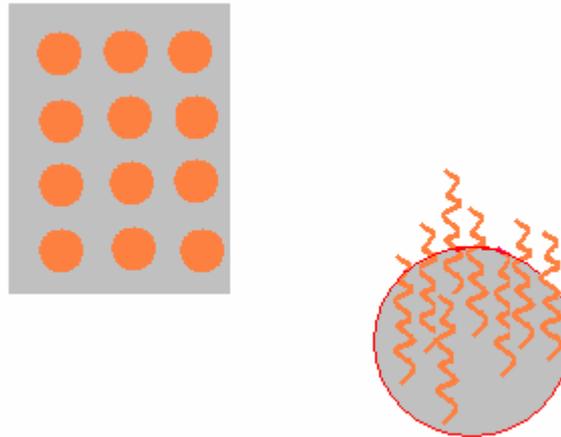


Figura 9. Esquema de un sector de un microarreglo hipotético y un spot ampliado

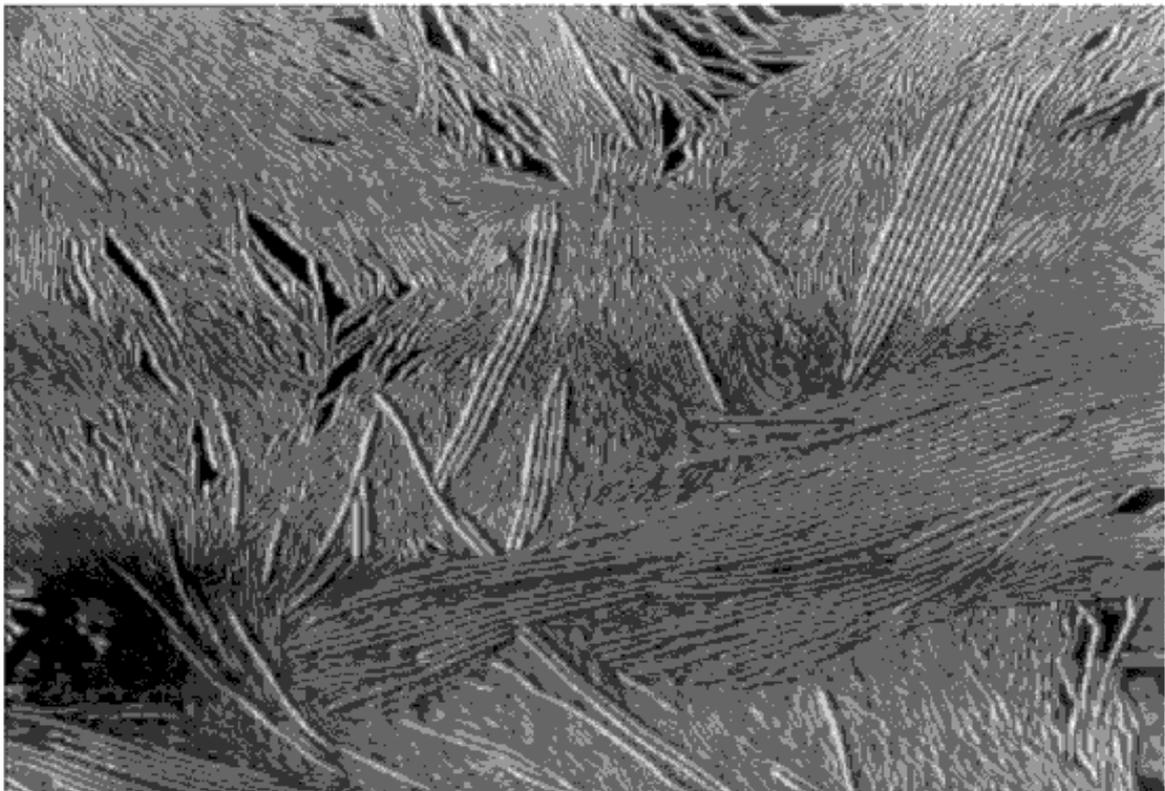


Figura 10: imagen obtenida mediante un microscopio electrónico de **un segmento** de un **spot de un microarreglo** –las hebras son las **moléculas de DNA** depositadas figura tomada de (**Duggan et al., Nature Genetics 21: 10-14, 1999**)

De acuerdo con el proceso de construcción, los microarreglos pueden clasificarse en dos grandes grupos:

- **Delivery: Microarrays de cDNA:** los probes son cADN (300-3000 bases) obtenidos por PCR de librerías -custom libraries- y espoteados sobre un portaobjetos de vidrio mediante un robot. **Oligos largos** Probes sintetizados aparte (off-line) (con técnicas como el PCR-polymerase chain reaction-, BAC - bacterial artificial chromosome- , phosphoramidite synthesis) y luego fijados al

soporte sólido (spotted DNA microarrays) por impresión de contacto.- Long-oligo spotted arrays: los probes tienen longitud uniforme (60-90 bases), espoteados como los de cDNA

- **Síntesis:** Se construyen los spots base por base directamente sobre el microarreglo mediante un proceso (fotolitografía, ink-jet). De este procedimiento resultan cadenas cortas de oligonucleótidos (25, 60 bases).

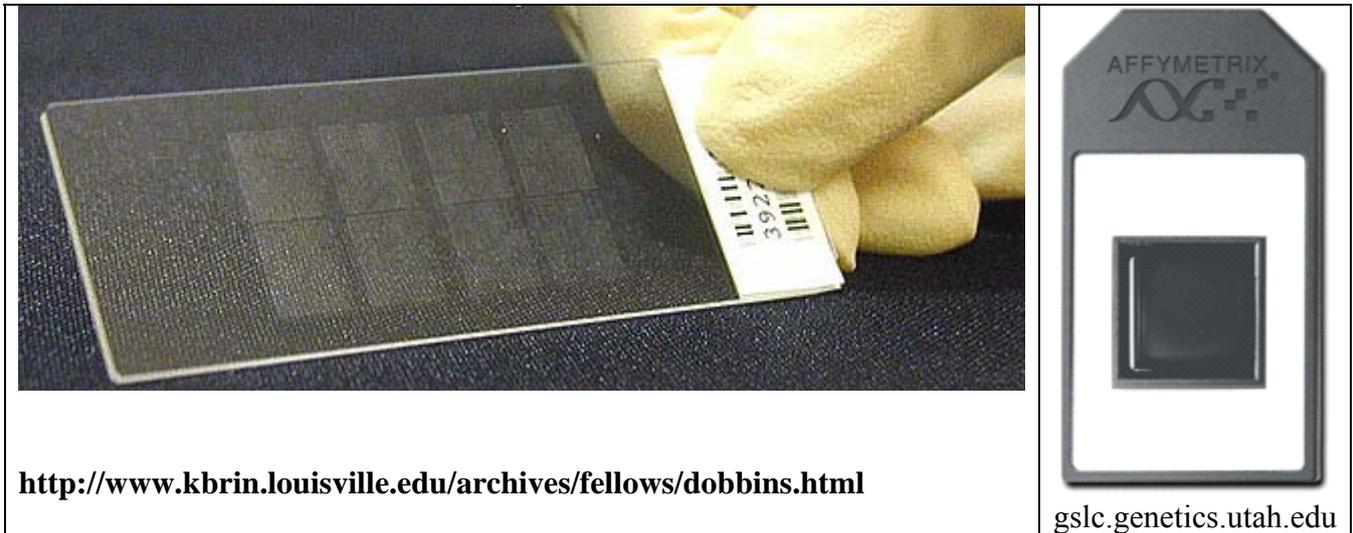


Figura 11. Dos tipos de microarreglos (vidrio, izq., silicio, der.)

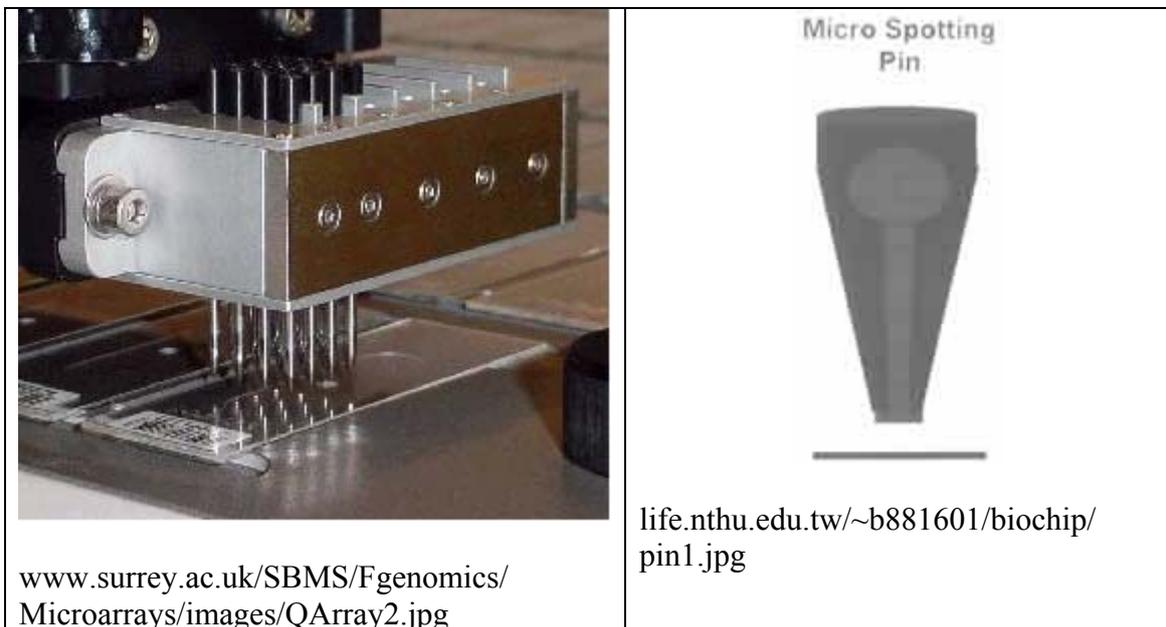


Figura 12. Cabezal de un robot y una aguja ampliada.

De acuerdo con el tipo de experimento, los microarrays se clasifican como.

Arregos de dos colores o dos canales -Two channel spotted arrays:

- Cada arreglo es hibridado con material proveniente de dos tipos de tejidos.

Un canal -Single channel arrays:

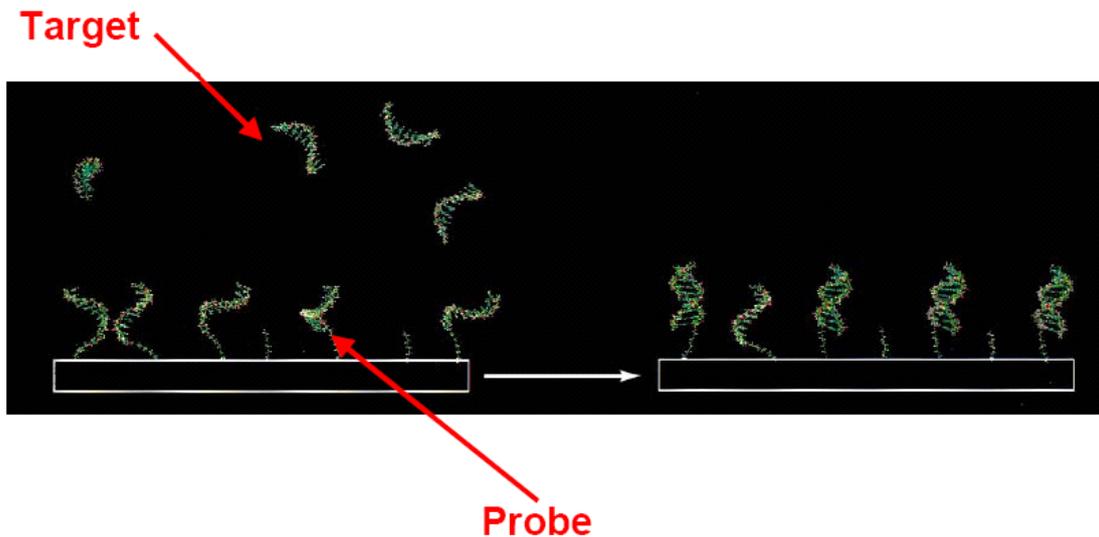
- Cada arreglo es hibridado con material proveniente de un tipo de tejido
Microarreglos de alta densidad: las sondas son oligos cortos (por ej, 25 bases Affymetrix, Nimblegen). Applied Biosystems AB1700

2.2 ¿Cómo actúan las sondas de un microarray?

El *principio* biológico de *complementaridad* por el que actúan estas sondas es el mismo que el que determina que el DNA en las células tenga una estructura de doble cadena. Establece que las secuencias de DNA o de RNA que contienen bases complementarias tienen una tendencia natural a pegarse:

. . . AAAAAGCTAGTCGATGCTAG . . .
. . . TTTTTCGATCAGCTACGATC . . .

Para cada secuencia determinada mRNA que interesa estudiar en un tejido (*target*, blanco, objetivo) se puede construir una sonda o *probe* utilizando el principio de complementaridad. La posición de la sonda nos indica la identidad del gen.

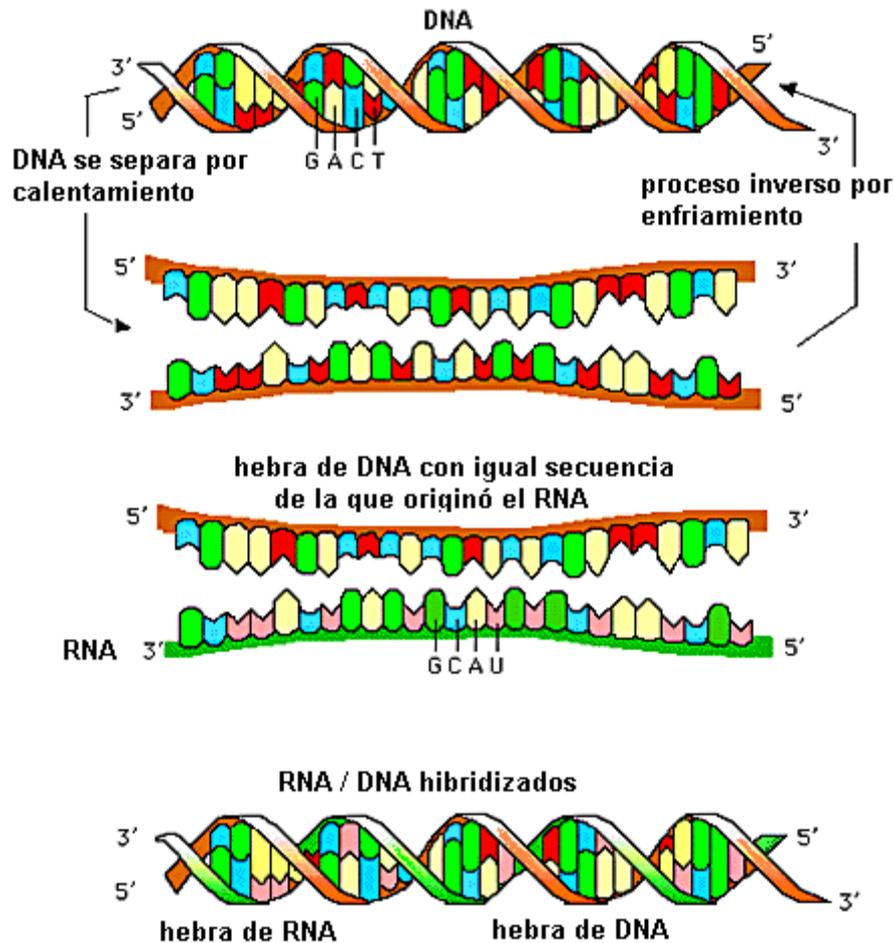


Bioconductor, ENAR03

Cada sonda del microarreglo actúa a modo de tubo de ensayo. Al poner una muestra correspondiente el material genético (mRNA) de un tejido en estudio en estas matrices, aquellas cadenas que tienen una *secuencia complementaria* a las del biochip se pegan por el principio de complementaridad, formando una doble cadena.

El proceso químico por el cual dos cadenas complementarias de ácido nucleico se enlazan (como si se subiera un cierre (zipper-up)) se denomina **hibridación**.

El proceso inverso, por el cual una doble cadena de ácidos nucleicos se separan, mediante calentamiento para destruir los puentes de hidrógeno se llama **desnaturalización**.



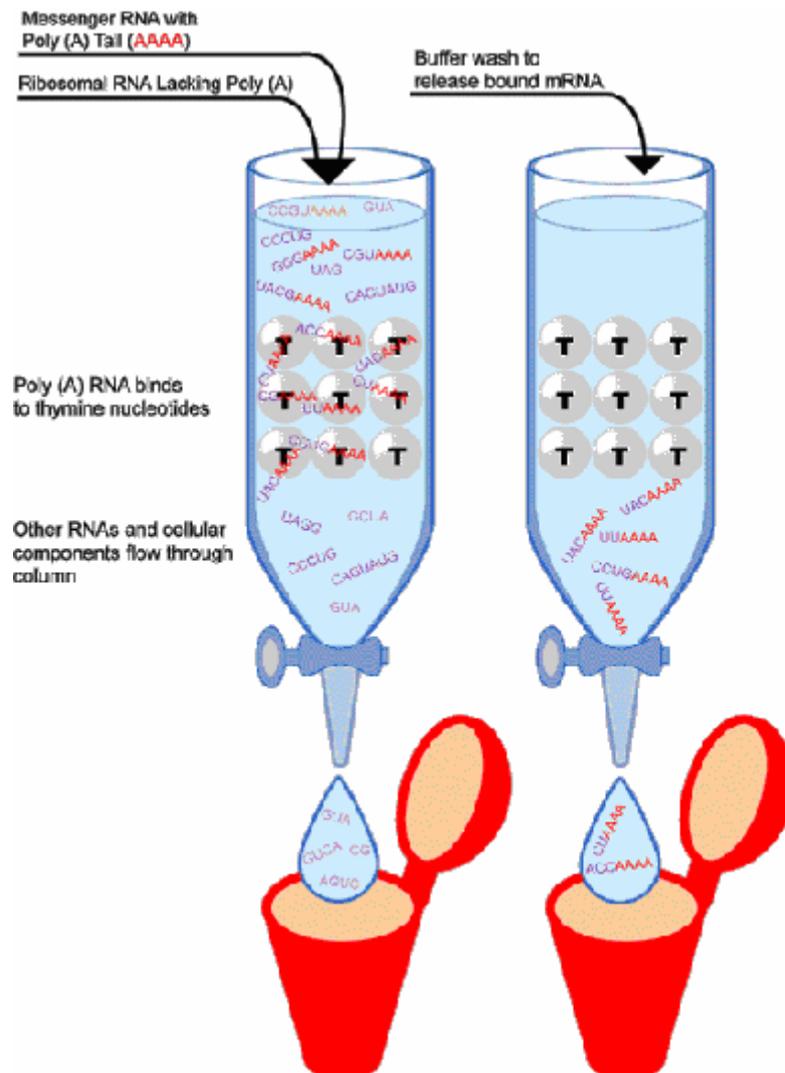
Hibridación y desnaturalización de ácido nucleico
www.accessexcellence.org/AB/GG/nucleic.html

Ambos procesos son importantes en los experimentos de microarreglos.

2.3 ¿Cómo se obtiene el material genético expresado de un tejido en estudio?

Aislación del mRNA maduro

Los métodos para aislar el mRNA aprovechan de la modificación natural que tiene el mRNA maduro. Como ya hemos visto parte del proceso normal de maduración del mRNA incluye un proceso de polyadenylation. Este proceso consiste en el agregado de unos 200 nucleótidos de adenina en una de las colas de la molécula del mRNA llamada cola poly(A). En el primer paso se rompe la célula y su contenido es expuesto sobre un soporte sólido que tiene un recubrimiento de hebras de nucleótidos T "synthetic beads coated with strings of" (thymine nucleotides). En este caso son bolitas de vidrio



http://www.ncbi.nlm.nih.gov/About/primer/genetics_molecular.html

Figura 13. Aislación del mRNA

Como las moléculas de timina y adenina tienen una gran afinidad para hibridar entre sí los mRNA con las colas de poly(A) quedan pegados a la superficie de las bolitas. Solamente es retenido el poly(A) RNA, porque queda inmobilizado al soporte sólido. El resto de los RNA's y demás material celular pasa a través de la columna. A la derecha se libera el poly(A) mRNA tratándolo con una solución especial (buffer solution) que rompe el enlace nucleótido timina - AAA. Puede así obtenerse el mRNA en un tubo para un proceso experimental posterior

Transcripción inversa

Una vez aislado, el mRNA purificado es convertido en DNA de una hebra mediante la enzima **transcriptasa reversa** (reverse transcriptase) y luego se sintetiza una cadena estable de DNA doble cadena mediante la enzima **DNA polimerasa** (DNA polymerase). Este DNA es llamado **DNA complementario (cDNA)** porque la primera de las hebras es complementaria al mRNA del cual fue producido.

¿Por qué se produce cDNA? Porque cDNA es un compuesto mucho más estable que el mRNA y porque al ser obtenido a partir de un mRNA en el cual las regiones no codificantes han sido removidas representa únicamente la secuencia de DNA expresada.

2.4 Experimento con microarreglos

El objetivo de un experimento de microarreglos es medir la cantidad de copias de cada gen en un tejido en estudio y compararla con la de un tejido control. Describimos a continuación las diferentes etapas que lo constituyen.

Extracción del tejido.

Extracción del mRNA: hemos visto este punto en la sección 2.3

Estandarización de la muestra: dilución de la muestra de mRNA en una cantidad especificada.

Spiking RNA: agregado de cantidades conocidas de genes no relacionados con la muestra. Estos niveles de expresión conocidos permitirán tener valores de referencia para los procedimientos posteriores de análisis de los datos. Si interesara comparar los niveles de expresión génica de una planta en dos condiciones experimentales diferentes se podrían agregar cantidades conocidas (por ej iguales) de uno o más genes humano a ambas muestras.

Etiquetado: labeling. Como el producto final del experimento de microarreglos es una imagen. El procedimiento estándar para lograr que los genes se vuelvan “visibles” es mediante el agregado de una tintura fluorescente. En el caso de microarreglos de dos canales se utilizan dos tinturas diferentes (Cy3 excitada con un láser “verde”, Cy5 excitada con un láser “rojo”) para distinguir la muestra tratada y control. Uno de los métodos más simples para realizar el etiquetado es el de transcripción inversa. Para que esta pueda realizarse además de la enzima es necesario agregar el material de construcción de la cadena, o sea los nucleótidos A, T, G, C. Pero en vez de agregar los nucleótidos C “limpios” se agregan con una molécula de tinte pegada

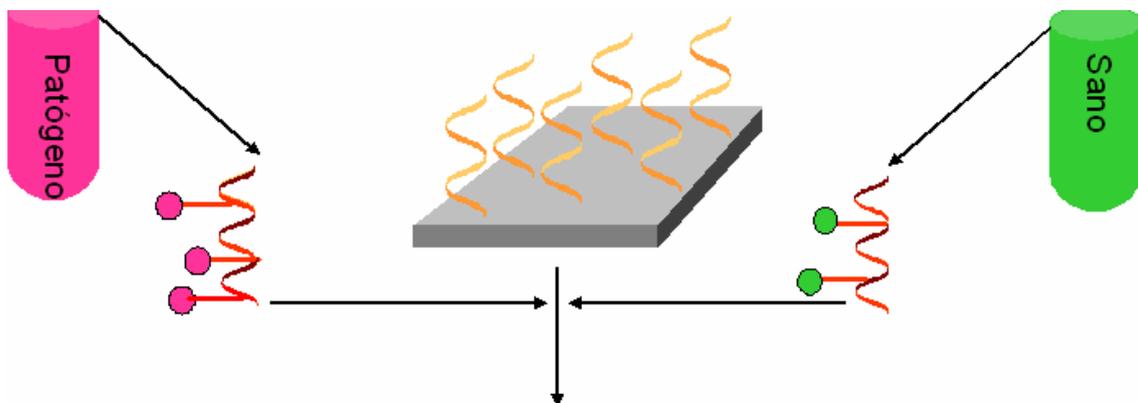


Figura 14: esquema del proceso de etiquetado, previo a la incorporación de la muestra sobre el microarreglo

Cada vez que una G requiera una C se pegará también el tinte. La cantidad de moléculas de tinte será proporcional a la cantidad de G's en el mRNA que es aproximadamente proporcional a la cantidad de copias transcriptas del gen y a su longitud.

Hibridación: Es un proceso complejo en el cual se forman dobles cadenas entre las sondas (probes) del microarreglo y las cadenas etiquetadas (target) de las muestras. Muchas son las condiciones que lo afectan: temperatura, humedad, concentración de sal, volumen de la solución del target, operador, etc. La mezcla de cDNA que contiene la (o las) muestra(s) se aplica sobre el arreglo con una pipeta. Se pone una cubierta hidrofóbica (hydrophobic) hasta que desaparecen las burbujas de aire y luego el vidrio es montado en un recipiente (hybridization chamber) fijo, en la oscuridad a una temperatura que está entre 45 y 65 °C dependiendo del tipo de arreglo que se utilice. La mayoría de las hibridaciones llevan de 12-24 hs. En este tiempo es cuando ocurre la hibridación: el cDNA que fue aplicado al vidrio se pega a las hebras complementarias del arreglo. La cantidad de moléculas que han hibridizado en cada punto determina la intensidad de la imagen escaneada dando un indicador de la cantidad de mRNA transcripto de ese gen en la muestra.

Lavado: Luego de la hibridación el vidrio es lavado para eliminar el exceso de la solución de hibridación y también para reducir la hibridación no específica. El objetivo es que solamente quede adherido al array las hebras complementarias al mismo.

Los dos últimos pasos pueden ser realizados manual o automáticamente.

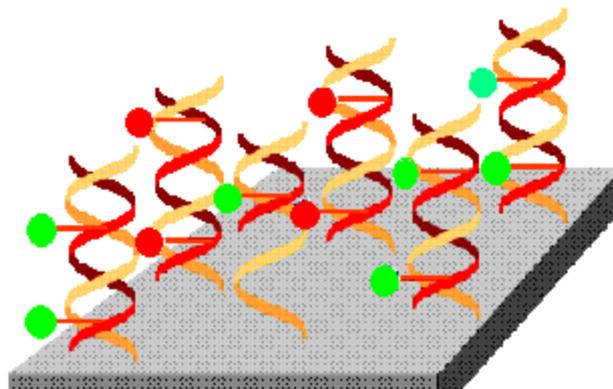


Figura 15: esquema del microarreglo en el que las sondas y el material objetivo (target) ha sido hibridizado

Obtención de la imagen: Veremos este punto en la sección 3.

2.5 Características especiales de los chips de alta densidad de oligos

Probes: cada probe está formado por una secuencia de 25 bases

Conjuntos de Probes (probe set): Cada gen está representado por un conjunto de probes. Para cada gen objetivo se seleccionan un conjunto de sectores que lo identifican en forma específica, en el pasado se utilizaron conjuntos de 20 probes en chips

humanos, actualmente hay 11 (Human GeneChips® HG-U133A). Para algún gen puede haber más de un probe set.

Un sólo canal: a cada chip se hibrida una muestra con un único tinte fluorescente

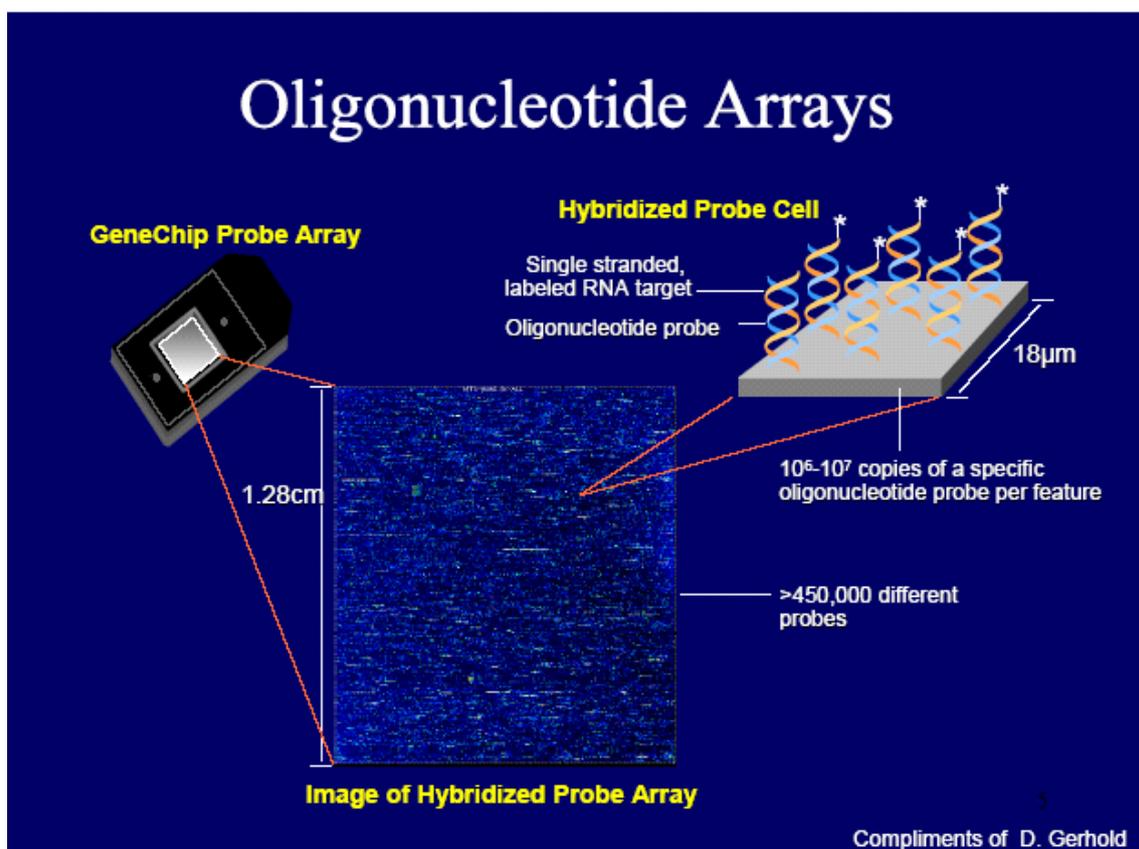
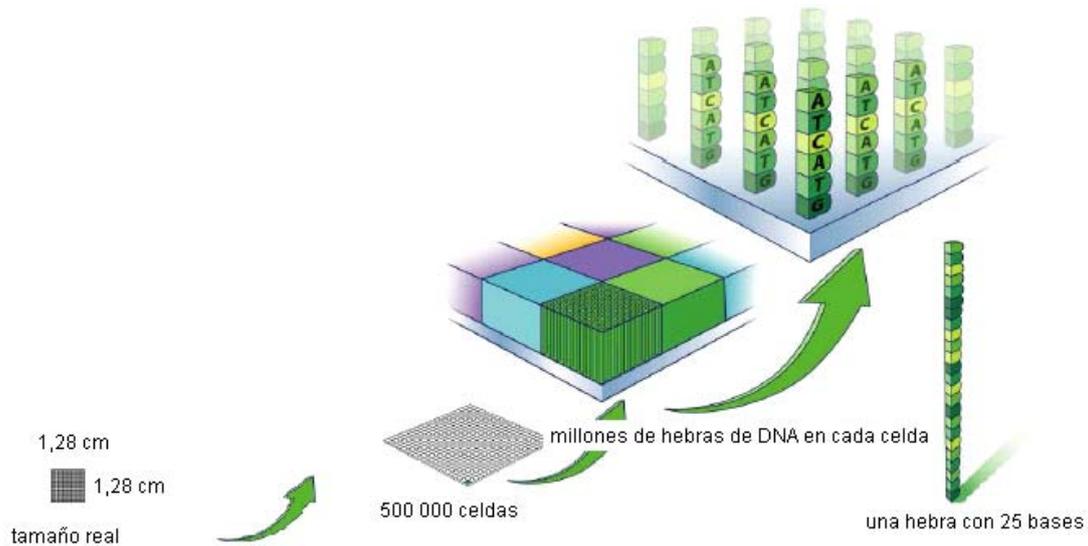


Figura 16a: Chips de alta densidad

PM.MM: Perfect Match , Miss Match

Perfect Match probe (PM) = probe de 25 bases que es perfectamente complementario a una región específica de un gen

Mis Match probe (MM) = probe de 25 bases que concuerda con un PM salvo en la base central, que es diferente (A → G, C → T, una transición de esa base)

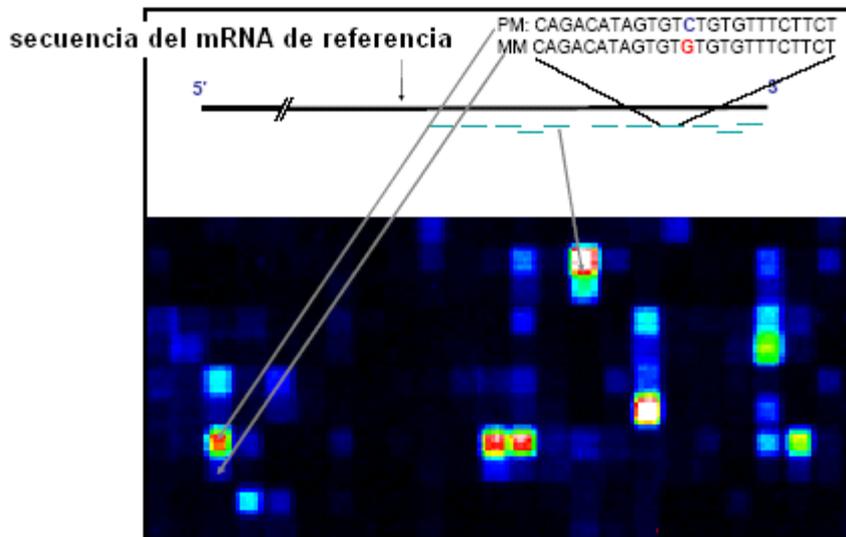
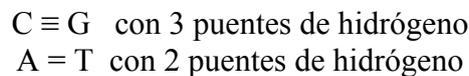


Figura 16b: representación de PM-MM

Los MMs fueron un intento de captar hibridación cruzada. Un microarreglo típico presenta un 30% de valores de MM superiores a los de PM. Affymetric incluía un MM por cada PM pero esto no seguirá así (Terry Speed Ago 2005)

2.6. Afinidad de pegado-binding affinity-

Hemos visto en la sección 1.2.1 que la unión entre las bases en una cadena de ADN se establece mediante puentes de hidrógeno:



de manera que la unión entre C y G es más fuerte que aquella entre A y T. Por lo tanto la fuerza general de unión entre dos cadenas de ADN depende del contenido de C+G.

Un factor importante en la afinidad entre el probe y el target es el contenido de C + G en el probe.

En general, los genes tendrán una longitud de unos cientos a unos miles de pares de bases y los probes serán más cortos en por lo menos un orden de magnitud. Esto resulta en parte por el costo en el proceso de construcción del microarreglo. La elección de las secuencias a elegir requieren que sean únicas al gen de interés (specific binding) pero suficientemente corta.

En el caso de los chips de Affymetrix la longitud de cada probe es de 25 bases.

Diferentes probes para el mismo gen tienen afinidades de pegado diferente.

Es difícil establecer si el gen A le gana al gen B. Lo que sí es posible es comparar el nivel de expresión de un mismo gen entre dos condiciones diferentes. Los microarreglos producen medidas relativas de los niveles de expresión.

Las afinidades son desconocidas. Para cubrirse de los problemas que puede tener un probe específico se utilizan varios probes para cada gen. Se desconoce cuál es la cantidad óptima. Sucesivas generaciones de chips han utilizado 20, 16, y 11 probes por gen que interesa interrogar.

Hay más dificultades en la elección de los probes:

- algunos genes son cortos, múltiples subsecuencias se van a superponer.
- los genes tienen una orientación y la degradación del ARN comienza preferentemente en una de los extremos (3' bias).
- el gen puede no ser lo que pensamos, las bases de datos están evolucionando.
- los probes pueden tener "hibridación cruzada" es decir que se pegan a targets equivocados.

La superposición no es demasiado grave. el problema de la orientación puede ser controlado eligiendo los probes más cerca de uno de los extremos. La hibridación cruzada se intenta controlar utilizando pares PM-MM

PM: GCTAGTCGATGCTAGCTTACTAGTC
MM: GCTAGTCGATGCAAGCTTACTAGTC

Lectura recomendada

Improving comparability between microarray probe signals by thermodynamic intensity correction. Georg M. Bruun, Rasmus Wernersson, Agnieszka S. Juncker, Hanni Willenbrock and Henrik Bjørn Nielsen, Nucleic Acids Research, 2007, Vol. 35, No. 7 e48
<http://nar.oxfordjournals.org/cgi/content/full/35/7/e48>

2.7 Algunas consideraciones sobre el diseño de los experimentos

Cualquier experimento de microarreglos involucra el diseño del microarreglo y el diseño de la muestra.

- **Diseño del arreglo:** esto es decidir qué sondas y donde, serán impresas al sustrato sólido.

- **Diseño de las muestras que se podrán sobre el arreglo:** decidir cómo deben prepararse las muestras de mRNA para la hibridación, cómo deben ser etiquetadas, naturaleza y cantidad de replicaciones a realizarse.

La elección de los DNA probes que serán impresos en el arreglo está determinada por el tipo de genes cuyos niveles de expresión desea medir la/el bióloga/o ó por las bibliotecas de cDNA (colecciones de clones de cDNA) accesibles a los investigadores.

Con los arreglos que se sintetizan in-situ, (oligo-microarreglos de alta densidad) en general esta decisión la toma la compañía fabricante (arreglos estándar) pero también existe la posibilidad de solicitar arreglos específicos (custom arrays). Muchos investigadores también compran vidrios con cDNA pre-espoteados. En el caso de los oligos cortos (25 bases) o de los oligos largos (60-75) bases la determinación de las secuencias a ser fijadas al sustrato es una cuestión importante y especializada de bioinformática (puede consultarse <http://www.affymetrix.com/technology/design/index.affx>, <http://www.genelink.com/Literature/ps/CAT-OLIGO.pdf>).

Los arreglos, además de contener los probes de interés, contienen spots de control:

- **controles negativos:** spots en blanco, spots impresos con la solución (buffer solution).
- **controles de nivel:** spots con cDNA de especies muy diferentes (por ej de bacterias cuando se están estudiando mamíferos) que serán agregados a las muestras (*spiked in*) en cantidades pre especificadas.
- **controles positivos:** “housekeeping genes”, son genes que están expresados a niveles semejantes en las muestras.

Algunos spots se incluyen con el objetivo de evaluar si la hibridación fue un éxito o un fracaso. Otros para facilitar las etapas de normalización (veremos esto más adelante) que se realizan para controlar las diferentes fuentes de sesgo de los experimentos o para evaluar la calidad de los resultados. Otro aspecto relacionado con el diseño del microarray es la *replicación*.

2.6.1 Spots duplicados

Es habitual incluir en el arreglo spots *duplicados* (una o más veces). Muchas veces estos son adyacentes. Permiten estimar la variabilidad de la señal, sin embargo esta estimación será en general menor que la observada entre distintos arreglos “idénticos” para un mismo spot.

2.6.2 Muestras replicadas

Replicaciones técnicas

Se denominan replicaciones técnicas cuando se realizan hibridaciones replicadas utilizando un mRNA objetivo (target), correspondiente a *una misma extracción biológica*. Generalmente la denominación replicaciones técnicas supone que la muestra de mRNA ha sido etiquetada en forma independiente para cada hibridación.

Inicialmente los laboratorios realizaban una replicación técnica más restringida, separando el material para la hibridación a partir de una única extracción y etiquetado.

Replicaciones biológicas

Las replicaciones biológicas de *tipo I* se refieren a los arreglos replicados cuyos targets fueron obtenidos de *diferentes muestras biológicas* de una línea celular (cell line) o de un mismo tejido (por ej. sangre de un mismo paciente).

Se denomina replicación biológica de *tipo II*, cuando los el target de los arreglos replicados provienen del *mismo tejido* pero de *diferentes individuos* de la misma especie o diferentes versiones de una línea celular. Esta forma de replicación biológica involucra un mayor grado de variabilidad en las mediciones.

2.7 Fuentes de sesgo debido a la secuencia

Describimos a continuación únicamente las fuentes de sesgo intrínsecas a un experimento de microarreglos debido a las diferencias entre las secuencias

- Las secuencias que tienen más guanina (G) aparecerán más brillantes ante la detección de la fluorescencia del microarreglo.
- Los pares C≡G tienen una afinidad más fuerte de hibridación que los pares A=T. Esto se traduce en diferentes temperaturas y tiempos de hibridación óptimos de acuerdo con la secuencia y su longitud.
- Diferentes moléculas de mRNA, con diferente secuencia y diferente longitud, tienen grados de eficiencia variable en el proceso de transcripción reversa.

Los sesgos dependientes de la secuencia impiden que sean válidas las comparaciones de las intensidades de distintos genes de un mismo arreglo. En cambio sí pueden compararse las intensidades de una secuencia determinada a través de diferentes arreglos.