



19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA

11 Número de publicación: **2 331 737**

51 Int. Cl.:
G06F 19/00 (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

96 Número de solicitud europea: **05077415 .7**

96 Fecha de presentación : **25.05.2000**

97 Número de publicación de la solicitud: **1630709**

97 Fecha de publicación de la solicitud: **01.03.2006**

54 Título: **Análisis matemático para la estimación de cambios en el nivel de expresión génica.**

30 Prioridad: **25.05.1999 US 135853 P**
24.05.2000 US 577634

45 Fecha de publicación de la mención BOPI:
14.01.2010

45 Fecha de la publicación del folleto de la patente:
14.01.2010

73 Titular/es: **Aventis Pharmaceuticals Inc.**
300 Somerset Corporate Boulevard
Bridgewater, New Jersey 08807, US

72 Inventor/es: **Theilhaber, Joachim;**
Bushnell, Steven y
Fuchs, Rainer

74 Agente: **Elzaburu Márquez, Alberto**

ES 2 331 737 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín europeo de patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre concesión de Patentes Europeas).

DESCRIPCIÓN

Análisis matemático para la estimación de cambios en el nivel de expresión génica.

5 **Campo de la invención**

La presente invención se refiere a un análisis matemático para la estimación cuantitativa del nivel de expresión génica diferencial. Más específicamente, la presente invención se refiere a la derivación matemática de una distribución *a posteriori* de todos los factores de cambio del nivel de expresión génica que pueden inferirse de las mediciones experimentales dadas.

Las células dependen de sus numerosos componentes proteicos para una amplia variedad de funciones. Estas funciones incluyen, p.ej., la producción de energía, la biosíntesis de todas las macromoléculas que la componen, el mantenimiento de la arquitectura celular, la capacidad de actuar frente a estímulos intra- y extracelulares, y similares. Cada célula dentro de un organismo contiene en ella la información necesaria para producir el repertorio de proteínas que ese organismo puede expresar. Esta información se almacena como genes en el genoma del organismo. El número de genes humanos únicos se estima que es de 30.000 a 100.000.

Para una célula dada, sólo una porción de la serie de genes se expresa en forma de proteína. Lo más probable es que algunas proteínas estén presentes en todas las células (esto es, se expresan de manera ubicua) porque cumplen función(es) biológica(s) que se requieren en todos los tipos de célula, y pueden parecer como proteínas “gobernantes”. Por el contrario, otras proteínas cumplen funciones especializadas que se requieren sólo en tipos celulares particulares. Por ejemplo, las células musculares contienen proteínas especializadas que forman las densas fibras contráctiles de un músculo. Dado que una gran parte de la funcionalidad específica de una célula está determinada por los genes que está expresando, es lógico que la transcripción, el primer paso en el proceso de convertir la información genética almacenada en el genoma de un organismo en proteína, estuviera altamente regulada por la red de control que coordina y dirige la actividad celular.

La regulación de la expresión génica se observa fácilmente en estudios que examinan actividades evidentes en células que se configuran a sí mismas para una función en particular (p.ej., la especialización en una célula muscular) o estado (p.ej., multiplicación activa o inactividad). Por consiguiente, según las células alteran su situación, puede observarse la transcripción coordinada de la(s) proteína(s) que se requieren para este “estado” biológico/fisiológico molecular. Este conocimiento global, altamente detallado, del estado transcripcional de la célula proporciona información sobre la situación de la célula, así como sobre el/los sistema(s) biológico(s) que controlan esta situación. Por ejemplo, el conocimiento de cuándo y en qué tipos de células se expresa el producto proteico de un gen de función desconocida proporcionaría pistas útiles en cuanto a la función probable de ese gen. La determinación de los patrones de expresión génica en células normales podría proporcionar un conocimiento detallado de la forma en la que el sistema de control consigue la activación y desactivación altamente coordinadas requeridas para el desarrollo y diferenciación de un organismo maduro a partir de una única célula huevo fertilizada. La comparación de los patrones de expresión génica en células normales y patológicas podría proporcionar “huellas” de diagnóstico útiles y ayudar a identificar funciones aberrantes que serían dianas razonables para una intervención terapéutica.

Lamentablemente, la capacidad para llevar a cabo estudios en los que se determine el estado transcripcional de un gran número de genes ha estado, hasta hace poco, inhibida por limitaciones en la capacidad para examinar las células con respecto a la presencia y abundancia de un gran número de productos de transcripción génica en un único experimento. Una limitación puede estar en el pequeño número de genes identificados. En el caso de los humanos, sólo unos pocos miles de proteínas codificadas en el genoma humano se han purificado físicamente y caracterizado cuantitativamente hasta algún punto. Otra limitación puede estar en la forma de los análisis de transcripción.

Dos abordajes con avances tecnológicos recientes han ayudado a los análisis de transcripción génica. La clonación de moléculas derivadas de productos de transcripción de tipo mRNA en tejidos particulares, y a continuación la aplicación de la secuenciación de alto rendimiento para los extremos de ADN de los miembros de estas genotecas ha proporcionado un catálogo de etiquetas de secuencias expresadas (ESTs, del inglés “expressed sequence tags”). Véase, p.ej., Boguski y Schuler, *Nat. Genetics* 10: 369-370 (1995). Estas “secuencias firma” pueden proporcionar identificadores inequívocos para una gran cohorte de genes.

Además, los clones de los que derivaron estas secuencias proporcionan reactivos analíticos que pueden usarse en la cuantificación de productos de transcripción de muestras biológicas. Los polímeros de ácido nucleico, ADN y ARN, se sintetizan biológicamente en una reacción de copia en la que un polímero sirve como molde para la síntesis de una cadena opuesta, que se denomina su complementaria. Después de la separación de las cadenas una de la otra (esto es, la desnaturalización), puede inducirse que estas cadenas se apareen, bastante específicamente, con otras cadenas de ácido nucleico que poseen una secuencia complementaria, en un proceso denominado hibridación. Esta unión específica puede ser la base de los procedimientos analíticos para medir la cantidad de una especie en particular de ácido nucleico, tal como el mRNA que especifica un producto génico de una proteína en particular.

Un segundo avance implica la tecnología de micromatrices/microensayos. Este es un procedimiento basado en la hibridación que permite la cuantificación simultánea de muchas especies de ácido nucleico. Véase, p.ej., DeRisi *et al.*, *Nat. Genetics* 14: 457-460 (1996), Schena *et al.*, *Proc. Natl. Acad. Sci. USA* 93: 10614-10619 (1996). Esta técnica

combina la colocación robótica (esto es, la aplicación de manchas (“spotting”)) de pequeñas cantidades de especies de ácido nucleico puras individuales sobre una superficie de vidrio, la hibridación de esta matriz con múltiples ácidos nucleicos marcados con fluorescencia, y la detección y cuantificación de los híbridos marcados con fluorescencia resultantes con, por ejemplo, un microscopio confocal de barrido. Cuando se usa para detectar productos de transcripción, un producto de transcripción de tipo ARN particular (esto es, un mARN) puede copiarse para convertirlo en ADN (esto es, un cADN) y esta forma copiada del producto de transcripción se inmoviliza subsiguientemente sobre, por ejemplo, una superficie de vidrio.

Un problema en el análisis de los datos de expresión génica es la estimación del factor de cambio global en el nivel de expresión de un gen en un experimento con relación a su expresión en otro experimento. Dadas estas dos mediciones brutas del factor de cambio en el nivel de expresión génica, el enfoque más simple, tal como se ha utilizado en metodologías anteriores, ha sido tomar la relación aritmética de los valores como un valor estimado del factor de cambio global. Mientras que para señales muy fuertes esto conduce a un valor estimado significativo del factor de cambio en las concentraciones de mARN subyacentes, para señales más débiles los resultados son mucho más ambiguos debido a la contaminación por el “ruido” que es autóctono para el sistema experimental particular utilizado. Otra tecnología anteriormente utilizada para la estimación del factor de cambio en el nivel de expresión génica se basa en intensidades de señal diferentes (p.ej., el chip Affymetrix®). Sin embargo, los valores asignados a los niveles de expresión usando la metodología anteriormente mencionada pueden ser negativos, conduciendo así a la situación delicada de relaciones de expresión génica negativas o indefinidas.

El documento WO-A-99/54724 describe un procedimiento analítico para discriminar datos adquiridos a partir de muestras con distribuciones que solapan, y para mejorar y evaluar la validez estadística de la señal de hibridación en matrices de ensayos. El procedimiento incluye un método para convolucionar datos en dos o más funciones de densidad de probabilidad discretas que representan señal y no señal, flujos discretos u otras variables independientes convolucionadas. El sistema usa las funciones de densidad de probabilidad para asignar las señales de hibridación, objetivamente, a una de las distribuciones modelo. Los procedimientos posteriores evalúan la viabilidad inherente a las matrices, y usan esta variación evaluada para establecer escalas de fiabilidad y límites de confianza para matrices de hibridación completas, y para ensayos de hibridación discretos dentro de las matrices.

De acuerdo con un aspecto de la presente invención, se proporciona un método para calcular indicios de diferencias en la expresión génica en una pluralidad de hibridaciones de matrices como se reivindica en la reivindicación 1 más adelante.

La invención también comprende un producto que es un programa de ordenador para calcular tales indicios según se reivindica en la reivindicación 7 más adelante.

Por tanto, la presente invención proporciona una metodología basada en las matemáticas altamente precisa y reproducible para cuantificar los niveles de expresión génica diferencial a partir de protocolos de micromatrices.

Los métodos de la presente invención pueden usarse para calcular diferencias en el nivel de expresión génica en una pluralidad de matrices de genes. Los métodos implican la definición del ruido experimental asociado con la intensidad de la señal de hibridación para cada gen en las matrices. El ruido experimental es las variaciones en los niveles observados en los chips u otras micromatrices más que el ruido biológico, que es la variación del nivel de expresión visto en sistemas biológicos. La detección de genes se basa a menudo, pero no siempre, en la fluorescencia. Se han usado otros sistemas de detección que pueden adaptarse aquí. Tales sistemas incluyen marcadores luminiscentes o radiactivos, etiquetas biotiniladas, hapténadas, u otras etiquetas químicas que permiten una fácil detección de las sondas marcadas.

Para una descripción matemática, véase la Sección 1 - Formulación del Modelo de Ruido más abajo. Se supone que el ruido es gaussiano y se aplica el Teorema de Bayes. El término de ruido experimental definido, sigma, se usa seguidamente para definir una función de distribución de probabilidad (“pdf”, del inglés “probability distribution function”) analítica (esto es, analítica en el sentido matemático, que significa que es una función continua) que describe los valores de distribución de la intensidad para cada gen. Estas pdfs se usan para derivar una pdf conjunta analítica que describe las relaciones posibles o factores de cambio para cualquier gen o producto génico expresado de manera diferencial en la(s) matriz(ces). Las pdfs conjuntas se aplican usando intensidades y valores de ruido derivados experimentalmente de los genes en la(s) matriz(ces) (1) para estimar los factores de cambio en la concentración de los productos de transcripción génica, (2) para usar la jpdf (del inglés “joint probability distribution function”, función de distribución de probabilidad conjunta) para establecer los límites de confianza en el factor de cambio dados unos intervalos de confianza específicos, y (3) para derivar un valor p, o medidor de la calidad (la probabilidad de que un factor de cambio pudiera ser menor que 1, cuando el valor estimado es mayor que 1, o, la probabilidad de que el factor de cambio sea mayor que 1, cuando el valor estimado es menor que 1), asociado con el valor estimado del factor de cambio. El factor de cambio estimado determinado por los métodos de la presente invención representa la diferencia en el nivel de expresión génica observado. La varianza total (esto es, el ruido) puede aún ser alta incluso a medida que la concentración del producto de transcripción se aproxima a cero. Los métodos de la presente invención usan una fórmula matemática para describir una distribución estadística *a posteriori* de todos los niveles de expresión génica que pueden derivarse de las mediciones obtenidas de los niveles de expresión génica en uno o más tipos de células o tejidos representados en la(s) matriz(ces).

Las micromatrices son una matriz ordenada de moléculas de ADN de cadena doble o monocatenarias colocadas en un material de soporte en una organización espacialmente separada. En contraste con las “macromatrices” en filtro, que son típicamente láminas grandes de nitrocelulosa, las micromatrices colocan el ADN en una organización empaquetada de manera más densa, de tal forma que pueden meterse hasta 10000 moléculas de ADN en una región típicamente de 1-4 centímetros cuadrados. Las micromatrices usan típicamente vidrio revestido como soporte sólido, en contraste con el material a base de nitrocelulosa de las matrices en filtro. Teniendo una matriz ordenada de muestras de ADN, la posición de cada muestra puede localizarse y relacionarse con la muestra original a partir de la cual se generó el ADN en la matriz. Se han descrito métodos y aparatos para preparar una micromatriz. Véase, p.ej., las Patentes de los Estados Unidos 5.445.934 y 5.800.992.

Las muestras de ADN en la micromatriz se hibridan con sondas de ARN o ADN que se han marcado con fluorescencia para identificar si la muestra con la sonda contiene una molécula que es similar o idéntica a la muestra de ADN en la micromatriz. En las condiciones apropiadas, las moléculas sonda se hibridan con una molécula de ADN en la micromatriz. Generalmente, las secuencias idénticas o casi idénticas forman híbridos productivos. La presencia de moléculas híbridas de ADN-sonda de ADN se detecta mediante un instrumento de detección de fluorescencia. Si la señal de hibridación es débil o inexistente en un sitio de ADN en particular, entonces la molécula de ADN o ARN correspondiente está ausente en la sonda. Los instrumentos de micromatrices actuales pueden hibridar hasta cuatro muestras con sondas fluorescentes diferentes de una vez. Con mejoras en la tecnología, pueden hibridarse más sondas de una vez.

Hasta hace poco, las hibridaciones de ADN se realizaban en filtros de nitrocelulosa. En contraste con las micromatrices donde el ADN se aplica como manchas directamente sobre la micromatriz, las matrices en filtro se generan aplicando manchas de colonias bacterianas sobre los filtros, colocando los filtros sobre un medio de crecimiento en agar, e incubando los filtros en condiciones que promueven el crecimiento de las colonias bacterianas. En ADN dentro de las colonias bacterianas se libera lisando la colonia y tratando los filtros para fijar el ADN al material del filtro. El procedimiento de generar una matriz bacteriana en un filtro puede llevar típicamente 2-4 días. Las micromatrices tienen varias ventajas frente a los métodos de matriz en filtro. Por ejemplo, los métodos en filtro generalmente presentan en la matriz colonias bacterianas en las que está contenido el cADN clonado. Las colonias deben crecerse durante varios días, lisarse para liberar el ADN y fijar el ADN sobre el filtro. La hibridación con las matrices en filtro de colonias es menos fiable debido a los restos bacterianos y la baja cantidad de ADN liberado de la colonia. Una segunda ventaja es que las iteraciones son más rápidas con las micromatrices que con los filtros. Esto se debe al tiempo necesario para crecer las colonias sobre los filtros y prepararlos para la siguiente tanda de hibridación. Por el contrario, el tratamiento con la sonda de una micromatriz subsiguiente puede comenzar menos de 24 h después de completar el análisis de una matriz. Otra ventaja de las micromatrices es la capacidad de usar sondas marcadas con fluorescencia. Esto asegura un método no radiactivo para detectar la hibridación. Por el contrario, la hibridación en filtro usa generalmente sondas marcadas con fósforo o azufre radiactivos. Las micromatrices pueden hibridarse con múltiples sondas simultáneamente. Por el contrario, las matrices en filtro sólo pueden hibridarse con una sonda cada vez. Una de las ventajas más importantes de las micromatrices es su reproducibilidad y sensibilidad de las señales de hibridación. Típicamente, las señales de hibridación son más elevadas y la sensibilidad es mayor en las micromatrices versus las matrices en filtro. Además, las matrices en filtro muestran a menudo señales de fondo falsas que no están relacionadas con la hibridación productiva entre la sonda y el ADN en el filtro.

Una vez que la muestra aleatoria de fragmentos de ácido nucleico se inmoviliza sobre una superficie sólida (p.ej., vidrio) en una micromatriz, la muestra aleatoria de fragmentos de ácido nucleico puede hibridarse seguidamente con una o más sondas marcadas complementarias a los genes o secuencias de interés. Generalmente, las sondas no hibridadas se eliminan. Las sondas marcadas se detectan seguidamente por métodos conocidos en la técnica (p.ej., microscopía confocal). Por ejemplo, las imágenes del portamuestras pueden analizarse con el software de análisis de imágenes Array Vision (Imaging Research) para el análisis para encontrar las manchas, la determinación del fondo localizado, la distribución de las intensidades de la señal en una mancha, y las relaciones de señales frente a ruidos. La evaluación estadística se realiza seguidamente como se describe más abajo.

La presente invención utiliza una metodología basada en las matemáticas para cuantificar el factor de cambio en los niveles de genes expresados de manera diferencial. Específicamente, la presente invención usa un enfoque deductivo simple, cimentado en un entorno bayesiano, para salvar la limitación basada en la heurística de las metodologías previas usadas en el análisis matemático de la expresión génica diferencial. La presente invención, más que buscar inmediatamente un valor estimado puntual del factor de cambio del nivel de expresión génica, deriva una fórmula matemática para la distribución *a posteriori* de todos los factores de cambio de la expresión génica diferencial que pueden inferirse a partir de las mediciones dadas. De esta distribución *a posteriori* puede obtenerse la siguiente información: (i) un estimador para el factor de cambio del nivel de expresión génica; (ii) límites de confianza para el factor de cambio, en cualquier nivel de confianza dado, y (iii) un valor P para evaluar la significación estadística del cambio. Una ventaja adicional de la presente invención es que los valores estimados del factor de cambio y límites de confianza pueden asignarse incluso a pares de señales donde ambas señales son cero o negativas, sin recurrir a límites heurísticos. Por consiguiente, el entorno matemático descrito en la presente memoria unifica la estimación para todas las señales en una muestra dada.

A continuación se da una descripción a modo de ejemplo y con referencia a los dibujos adjuntos de las realizaciones de la presente invención.

La Fig. 1 es una gráfica de línea que ilustra una distribución *a posteriori* del factor de cambio, Ecuación (16), para una serie de pares de mediciones (x_1, x_2). En todos los casos salvo en (0, 0), la relación de las mediciones es 4. La desviación estándar de ambos términos de ruido se mantiene constante a $\sigma_1 = \sigma_2 = 20$.

La Fig. 2 es una ilustración cualitativa de la derivación de la Ecuación (16), que explica el comportamiento de las distribuciones mostradas en la Fig. 1. Para cada par de señales (x_1, x_2), se dibuja una caja con límites $\pm \sigma_c$ alrededor de un punto en el plano. Se dibujan líneas desde el origen hasta el punto en la caja: la distribución de las pendientes de estas líneas es la distribución *a posteriori* de los factores de cambio. El Panel A representa la construcción para las señales (100, 400). El Panel B representa la construcción para las señales (5, 20).

La Fig. 3 es un trazado gráfico de las intensidades (x_1, x_2) en el plano (R, P), tal como se inducen por las Ecuaciones (30) y (33). Se muestran las líneas para x_1 y x_2 constantes. La línea oscura es el límite superior $P_u(R)$ del intervalo para P.

La Fig. 4 ilustra, en forma tabular, los resultados de los valores estimados R del factor de cambio, los intervalos de confianza del 68% (R_p, R_{p-1}), y los valores P para un factor de cambio positivo significativo, para todos los pares de mediciones de la Fig. 1.

La Fig. 5 ilustra, en forma tabular, los resultados de la realización de tres metodologías estadísticas utilizadas para la detección de 1000 genes con un factor de cambio real de $b = 3$ frente a un fondo de 4000 genes sin cambio ($b = 1$), basados en las simulaciones de Monte Carlo descritas en la Sección 6 de la presente memoria (esto es, la probabilidad *a priori* de que un gen sufra el factor de cambio 3 es $P_1 = 0,2$). Se presentan resultados específicos para un ratio de falsos positivos fijo $P(0 | p) = 0,3$.

La Fig. 6 es una dispersión de gráficos de dispersión en el plano (R, P) generados por: (1) 1000 genes de clase 0, la clase sin cambios (Panel A), y (2) 1000 genes de la clase 1 (Panel B).

La Fig. 7 es una gráfica de línea que representa una característica del operador del receptor (ROC, del inglés “receiver operating characteristic”) para el estadístico $t = P$.

La Fig. 8 es una gráfica de línea que representa una característica del operador del receptor (ROC) para el estadístico $t = -P$.

La Fig. 9 es una gráfica de línea que ilustra la sensibilidad comparada como una función de P_1 = fracción de genes que sufrieron un cambio ($b = 3$), para los estadísticos R y P. Se impone un ratio relativo de falsos positivos fijo $P(0 | p) = 0,3$ de principio a fin.

La Fig. 10 es un organigrama de un ordenador para implementar las características de la invención.

La Fig. 11 es un diagrama de flujo de un procedimiento según la invención.

Con referencia a la Fig. 10, un ordenador 102 incluye un procesador 104, una memoria 106, dispositivos de disco 108, un visualizador 110, un teclado 112, y un ratón 114. El procesador 104 puede ser una unidad central de proceso (CPU, del inglés “central processing unit”) de ordenador personal tal como un procesador Pentium® III fabricado por Intel® Corporation. La memoria 106 incluye la memoria de acceso aleatorio (RAM, del inglés “random access memory”) y la memoria de sólo lectura (ROM, del inglés “read-only memory”). Los dispositivos de disco 108 incluyen un dispositivo de disco duro y pueden incluir dispositivos de disquetes, un dispositivo de CD-ROM, y/o un dispositivo de compresor. El visualizador 110 es un tubo de rayos catódicos (CRT, del inglés “cathode-ray tube”), aunque son aceptables otras formas de visualizador, p.ej., visualizadores de cristal líquido (LCD, del inglés “liquid-crystal displays”) incluyendo visualizadores TFT. El teclado 112 y el ratón 114 proporcionan los mecanismos de entrada de datos para un usuario (no mostrados). Los componentes 104, 106, 108, 110, 112, y 114 están conectados por un bus 116. El ordenador 102 puede almacenar, p.ej., en la memoria 106, un código software que contenga las instrucciones para controlar el procesador 102 para realizar las funciones descritas más abajo.

1. Formulación del Modelo de Ruido

Las mediciones del nivel de expresión de un gen dado, en experimentos diferentes o réplicas, se basan en x , que puede escribirse:

$$x = Cn + \varepsilon, \quad (1)$$

donde n es la concentración física absoluta (molaridad) del mRNA del gen en solución, donde C es una constante de proporcionalidad que relaciona la molaridad con la intensidad registrada, y donde ε es un término de ruido. En las siguientes ecuaciones, no se busca la determinación de la concentración absoluta de mRNA y, por lo tanto, por simplicidad, el siguiente valor se establece como $C = 1$:

$$x = n + \varepsilon. \quad (2)$$

En la Ecuación (2) el término de ruido puede descomponerse en tres contribuciones separadas:

$$\varepsilon = \varepsilon_b + \varepsilon_c + \varepsilon_p, \quad (3)$$

donde ε_b es una variación que surge de fluctuaciones en la intensidad de fondo, donde ε_c es un término que surge de la hibridación cruzada de otros mARNs (específicas o no específicas), y donde ε_p es un término de “variación proporcional”, que surge de las variaciones de un chip a otro en la densidad de oligonucleótidos o cADN, y otros factores similares. Por ejemplo, con los chips Affymetrix®, donde la señal final x se obtiene haciendo la media de varias señales diferenciales, cada uno de los términos de ruido puede ser positivo o negativo, con una media de aproximadamente cero. Por consiguiente, los términos de fondo e hibridación cruzada pueden agruparse en un único término de ruido:

$$\varepsilon_c = \varepsilon_b + \varepsilon_c \quad (4)$$

y la media y la varianza del ruido total ε se escriben como:

$$\langle \varepsilon \rangle = 0, \quad (5)$$

$$\sigma_c^2 = \text{var}(\varepsilon_c) + \text{var}(\varepsilon_p) = \sigma_c^2 + (\alpha n)^2, \quad (6)$$

donde α es el parámetro de variación proporcional, de tal forma que $\text{var}(\varepsilon_p) = \sigma_p^2(\alpha n)^2$. El término de variación proporcional α es similar al coeficiente de variación c (inicialmente definido por Chen *et al.*, J. Biomed. Optics 2:364 (1997)). Otros términos de ruido intervienen también en este modelo, de manera que la varianza total del ruido puede continuar siendo grande incluso a medida que $n \rightarrow 0$. En las siguientes ecuaciones se supone que ε se distribuye normalmente.

Utilizando los chips Affymetrix® como un ejemplo, la varianza combinada del ruido de fondo y de la hibridación cruzada σ_c^2 se calcula tomando la varianza de las mediciones x para todos los genes señalizados como *ausentes* por el algoritmo de decisión de ausencia/presencia. El término proporcional se ha estimado comparando el cuartil más alto de intensidades en experimentos replicativos.

Para ilustrar la magnitud de los términos en la Ecuación(6) para un experimento basado en un chip Affymetrix® típico, el valor de la mediana de los niveles de expresión x es $\text{Med}(x) \approx 80$, con $\sigma_c \approx 25$ y $\alpha \approx 0,25$. Por lo tanto, la relación de señal mediana frente a ruido $\text{Med}(x)/\sigma_c$ es sólo aproximadamente 3. La desviación estándar del ruido de fondo solo es $\sigma_b \approx 3-4$, de manera que $\sigma_b \ll \sigma_c$ con el ruido de hibridación cruzada dominando el ruido de fondo por casi un orden de magnitud.

Debido a que la variable σ_c es relativamente grande, y α es pequeño, la Ecuación(6) puede simplificarse un poco escribiendo $n \approx x$ en el lado derecho de la ecuación:

$$\sigma_\varepsilon^2 = \sigma_c^2 + (\alpha x)^2, \quad (7)$$

de manera que no se necesita conocer la concentración subyacente de antemano para estimar cuantitativamente la varianza del ruido.

2. Distribución *A posteriori* de las Concentraciones

Mientras que la Ecuación(2) da la medición en términos de la concentración, la invención obtiene la concentración como una función de la medición. Esto puede formularse en términos probabilísticos, escribiendo el Teorema de Bayes para la variable n y x :

$$P(n|x) = \frac{P(n|x)P(n)}{P(x)} \quad (8)$$

En la Ecuación (8), $P(x | n)$ es la función de distribución de probabilidad (pdf) condicional para x , condicional sobre n , $P(n)$ es la distribución *a priori* de n (por lo tanto, que refleja el estado de conocimiento de n antes de la medición que se está tomando realmente), y $P(x)$, la pdf para x , es esencialmente un término de normalización. Por consiguiente, a partir de la Ecuación (2) y con la suposición de ruido gaussiano, puede derivarse la siguiente ecuación:

$$P(x | n) = \frac{1}{(2\pi\sigma_\epsilon^2)^{1/2}} \exp(-(x - n)^2 / 2\sigma_\epsilon^2) \quad (9)$$

en la que $\sigma_\epsilon = \sigma_c(n)$, Ecuación (6).

Para la distribución $P(n)$, como conocimiento *a priori*, se utiliza sólo el hecho de que la concentración es necesariamente no negativa:

$$P(n) = \begin{cases} 0 & n < 0 \\ \mu^{-1} n^\mu, & n \geq 0 \end{cases} \quad (10)$$

donde el límite $\mu \rightarrow 0$ rápidamente (esto es un mero artificio para conseguir una función de distribución paso a paso en el límite $\mu \rightarrow 0$, mientras que se mantiene $P(n)$ integrable en todo momento). Finalmente, $P(x)$ se obtiene por integración:

$$P(x) = \int_{-\infty}^{\infty} dn P(n) P(x | n). \quad (11)$$

Con el límite $\mu \rightarrow 0$, la Ecuación (8) puede reescribirse de la siguiente manera:

$$P(n | x) = \frac{P(n | x)}{\hat{P}(x)}, \quad n \geq 0, \quad (12)$$

donde $P(x | n)$ viene dado por la Ecuación (9) y donde el denominador es ahora:

$$\hat{P}(x) = \int_{-\infty}^{\infty} dn P(x | n). \quad (13)$$

La Ecuación (13) puede evaluarse fácilmente usando funciones de error. Más que explorar directamente las consecuencias de la Ecuación (12) en la estimación de las concentraciones, se utilizará más abajo para cuantificar la distribución de los factores de cambio.

3. Distribución *A posteriori* de los Factores de Cambio

Para un gen dado, puede suponerse que se desee evaluar el factor de cambio en el nivel de expresión génica entre dos experimentos dados (p.ej., los Experimentos 1 y 2). Por ejemplo, si las concentraciones de mRNA en los experimentos son n_1 y n_2 , respectivamente, entonces el factor de cambio R de la concentración en el Experimento 2 con relación al Experimento 1, viene dado por:

$$R = \frac{n_2}{n_1}. \quad (14)$$

Mientras que en la Ecuación (14) no hay acceso directo a n_1 y n_2 , la estimación de R en términos bayesianos puede formularse de manera inmediata escribiendo la distribución *a posteriori* de R como:

$$f_R(R | x_1, x_2) = \int_{-\infty}^{\infty} dn_1 \int_{-\infty}^{\infty} dn_2 \delta\left(\frac{n_2}{n_1} - R\right) P(n_1 | x_1) P(n_2 | x_2), \quad (15)$$

donde x_1 y x_2 son las mediciones de intensidad en los Experimentos 1 y 2, respectivamente, donde $\delta(\dots)$ se refiere a la función delta de Dirac, y donde $P(n | x)$ viene dado por la Ecuación (12) anterior.

Realizar la integración indicada en la Ecuación (15) es una tarea muy sencilla, si bien un poco tediosa. La función de distribución para R (esto es, omitiendo la dependencia explícita sobre x_1 y x_2 en $f_R(R | x_1, x_2)$) se obtiene por:

$$f_R(R) = \frac{C(x_1)C(x_2)}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x_1^2(R-R_0)}{2(\sigma_2^2 + R^2\sigma_1^2)}\right) I(x_1, x_2), \quad (16)$$

donde $\sigma_i^2 = \sigma_\varepsilon^2(x_i)$, $i = 1, 2$, con $a, (x)$ dado ahora por la Ecuación (7), con el término de normalización:

$$C(x) = \frac{2}{1 + \operatorname{erf}(x/\sqrt{2}\sigma_\varepsilon(x))}, \quad (17)$$

donde erf es la función de error (véase Abramowitz, M. y Stegun, I.A., p. 297 *Handbook of Mathematical Functions* (Dover, New York, 1972)), y con la definición:

$$I = \sigma_{12} \exp\left(-\frac{\alpha_{12}^2}{2\alpha_{12}^2}\right) + \alpha_{12} (2\pi\sigma_{12}^2)^{1/2} \frac{1}{2} \left(1 + \operatorname{erf}\left(\alpha_{12}/\sqrt{2}\sigma_{12}\right)\right), \quad (18)$$

donde:

$$\frac{1}{\sigma_{12}^2} = \frac{1}{\sigma_1^2} + \frac{R^2}{\sigma_2^2}, \quad (19)$$

$$\alpha_{12} = \left(\frac{x_1}{\sigma_1^2} + \frac{Rx_2}{\sigma_2^2}\right) / \left(\frac{1}{\sigma_1^2} + \frac{R^2}{\sigma_2^2}\right). \quad (20)$$

Aunque parece bastante compleja, la Ecuación (16) posee sólo dos límites simples que se discutirán más abajo a través de la utilización de dos escenarios.

3.1 Caso 1 - Altas Concentraciones

Si en ambos experimentos las concentraciones de ARN son grandes comparadas con la desviación estándar del ruido, con la consecuencia de que $x_i \gg \sigma_\varepsilon(x_i)$, $i = 1, 2$, R posee una distribución aproximadamente normal:

$$f_R(R) \approx \frac{1}{(2\pi\sigma_R^2)^{1/2}} \exp\left(-\frac{(R-R_0)^2}{2\sigma_R^2}\right), \quad (21)$$

En este límite, la media de R es justo la relación de las mediciones:

$$\langle R \rangle = R_0 = \frac{x_2}{x_1}. \quad (22)$$

Por consiguiente, la varianza σ_R^2 de R viene dada por:

$$\sigma_R^2 = \frac{\sigma_2^2 + x_2^2 \sigma_1^2 / x_2^2}{x_2^2} . \quad (23)$$

Usando la Ecuación (7), a su vez, puede obtenerse una aproximación simple para la desviación estándar de R:

$$\sigma_R = \sqrt{2\alpha} R_0 . \quad (24)$$

Por lo tanto, en el límite de alta concentración (esto es, Caso 1) la desviación estándar del factor de cambio real con respecto a la relación de la mediciones viene dada por una constante:

$$\frac{\sigma_R}{R_0} = \sqrt{2\alpha} . \quad (25)$$

Tras su análisis, la Ecuación (24) indica que sin importar cuánto de grandes sean la(s) señal(es), quedará una variación irreducible en la estimación del factor de cambio de orden $\sqrt{2\alpha} (\approx \pm 35\% \text{ para } \alpha = 0,25)$ del factor de cambio global que se va a medir.

3.2 Caso 2 - Concentraciones Muy Bajas

Si en ambos experimentos las concentraciones de ARN son tan bajas que $x_i \ll \sigma_\varepsilon(x_i)$, $i=1,2$, entonces la distribución toma la forma “universal” de:

$$f_R(R) \approx \frac{1}{\pi} \frac{1}{1+R^2} . \quad (26)$$

donde se supone que $\sigma_1 = \sigma_2$ por simplicidad.

En este límite, la distribución de R es completamente independiente de las concentraciones, la influencia de las cuales ha sido superada por el ruido. La Ecuación (26) define una denominada distribución de Cauchy (véase, p.ej., Keeping, E. S., *Introduction to Statistical Inference*, (Dover, New York, 1995)), que es muy ancha y no tiene una media finita. Una propiedad “patológica” autóctona de la distribución de Cauchy es que la media de muchas muestras independientes no converge, en modo alguno, a un único número, sino que queda distribuida según la Ecuación (26), sin tener en cuenta el número total de muestras utilizadas. Por el contrario, la mediana sobre la distribución de Cauchy es exactamente 1, y la mediana de las muestras converge a 1, sin ninguna de las patologías mencionadas anteriormente asociada con la mediana.

La función de distribución acumulativa de R viene dada por:

$$P(R \leq \rho) = \frac{2}{\pi} \tan^{-1} \rho . \quad (27)$$

Por ejemplo, los límites de confianza del 90% son (0,16, 6,3), lo que muestra que la distribución de la Ecuación (26) es muy ancha, puesto que estas cotas se obtendrán incluso cuando $R_0 = 1$, siempre que la señal sea suficientemente débil comparada con el ruido.

Finalmente, con la transformación $\mu = \log R$, la función de distribución de la Ecuación (26) se vuelve completamente simétrica:

$$f_u(u) = \frac{1}{\pi} \frac{1}{\cosh(u)} , \quad (28)$$

de manera que una representación logarítmica puede ser útil en algunos casos, aunque no se proseguirá más a fondo con los usos adicionales de la transformación en la presente memoria.

La Fig. 1 muestra la distribución *a posteriori* $f_R(R)$ para una serie de pares (x_1, x_2) , para una desviación estándar constante de ambos términos de ruido $\sigma_1 = \sigma_2 = 20$. En esta figura, la relación x_2/x_1 es siempre 4 (excepto para el caso donde ambas señales son 0), pero la relación de señal frente a ruido es muy variable. En los niveles de señal mas altos de todos, $(x_1, x_2) = (100, 400)$, y $f_R(R)$ forma un pico pronunciado alrededor de $R = 4$. Incluso en este límite, sin embargo, la inspección visual indica que el intervalo de confianza del 68% (correspondiente a un ancho de dos desviaciones estándar para una distribución normal) es aproximadamente (3, 5). Esto muestra que incluso cuando la relación de señal frente a ruido más baja es $100/20 = 5$, el factor de cambio real no puede inferirse a un valor mejor que $3 \leq R \leq 5$.

Con una relación de señal frente a ruido decreciente, la distribución $f_R(R)$ no sólo se ensancha, sino que su pico se desplaza hacia abajo. Por lo tanto, en la Fig. 1, para el par de mediciones (40,10), la mediana de la distribución es aproximadamente 2,2, con el máximo real teniendo lugar muy cerca de un valor de 1. Este ensanchamiento y desplazamiento de la función de distribución indica que para señales que se debilitan, la relación de las mediciones se convierte en una indicación cada vez menos fiable del factor de cambio real. Finalmente, en el límite de los valores medidos en que ambos son cero, (0, 0), la recuperación de la Ecuación (26) indica que la distribución es muy ancha, con mediana $R = 1$ y un pico en $R = 0$.

La Fig. 2 ilustra cualitativamente el comportamiento de la distribución $f_R(R)$, que se cuantificó previamente en la Fig. 1, a través de la utilización de la siguiente construcción: (i) para cada par de valores (x_1, x_2) , dibujar una caja en el plano alrededor del punto (x_1, x_2) con límites $\pm\sigma_\epsilon$ en cada dimensión, excluyendo las regiones que caen a lo largo de los ejes negativos y (ii) seguidamente dibujar una serie de líneas desde el origen a todos los puntos en la caja. La distribución de las pendientes de estas líneas representa la distribución *a posteriori* $f_R(R)$.

4. Estimación Bayesiana de los Factores de Cambio

La estimación bayesiana del factor de cambio R puede realizarse basándose en la Ecuación (16) y el conocimiento de las mediciones x_1 y x_2 . Inicialmente, la función de distribución acumulativa se define:

$$F(R') = P(R \leq R') = \int_0^{R'} f_R(R) dR. \quad (29)$$

$F(R)$ se evalúa preferiblemente usando integración numérica. Basándose en los valores numéricos de $F(R)$, puede obtenerse fácilmente la siguiente información.

4.1 Estimador del factor de cambio R

Se eligió el estimador de la mediana como un estimador R para el factor de cambio:

$$\hat{R} = Med(R), \quad (30)$$

esto es, el valor de R para el que $F(R) = 1/2$.

Son posibles otros estimadores, por ejemplo MAP (del inglés "Maximum *a posteriori* Probability", Probabilidad Máxima *a posteriori*) o los estimadores de la media. Véase, p.ej., Van Trees, H. L., *Detection, Estimation and Modulation Theory*, Part I (John Wiley and Sons), New York, 1998. Sin embargo, el estimador de la media no se usa aquí, puesto que $f_R(R)$ no tiene una media finita (esto es, tendrá siempre una "cola" con dependencia $1/R^2$, incluso en el límite casi normal de la Ecuación (21)). Por lo tanto, el estimador de la mediana tiene la doble ventaja de la solidez y la simetría bajo la transformación $(R \rightarrow 1/R)$, y es el que se adopta en la presente memoria. Formalmente, el estimador de la mediana es uno que reduce, p.ej. minimiza, el valor absoluto del término de error (valor estimado real). Véase, p.ej., Van Trees, H. L., *Detection, Estimation and Modulation Theory*, Part I (John Wiley and Sons), New York, 1998.

4.2 Límites de Confianza R_p y R_{1-p}

Dado $p < 1$, los límites de confianza R_p y R_{1-p} se definen como los valores de los percentiles correspondientes:

$$F(R_p) = p, \quad (31)$$

$$F(R_{1-p}) = 1 - p, \quad (32)$$

4.3 Valor P para la significación del cambio

La hipótesis $R > 1$ ("un factor de cambio positivo significativo tuvo lugar en el experimento 2 con relación al 1") puede probarse evaluando la probabilidad de la hipótesis complementaria, $R \leq 1$ y definiendo ésta como el valor P de la hipótesis para el cambio significativo. Esto se representa simplemente como:

$$P = F(R = 1). \quad (33)$$

Los resultados para todos los pares de mediciones discutidos en relación con la Fig. 1 se ilustran también en la Fig. 4 más abajo, con límites de confianza determinados por $p = 0,16$. Obsérvese que tener el valor P proporciona un poderoso criterio de selección para retener sólo aquellos pares de mediciones considerados significativos. Por lo tanto, aunque todas las relaciones de las mediciones mostradas en la Fig. 4 son iguales a 4 (excepto por supuesto para (0, 0)), sólo las tres primeras entradas ((100, 400), (50, 200), (25, 100)) se encuentra que indican un cambio significativo (esto es, en el nivel de confianza 0,05). A su vez, para cada una de estas entradas tabuladas mencionadas anteriormente, se conocen los límites de confianza para el factor de cambio. Por lo tanto, para el par de mediciones (25, 100), al valor estimado de $R = 3,6$ le siguen entre paréntesis los valores (2,0, 8,8), lo que muestra que en este ejemplo el factor de cambio no puede determinarse con más precisión que este intervalo señalado (esto es, los factores de cambio reales tan pequeños como 2 y tan grandes como 8,8 son consistentes con los datos).

5. Trazado de los Pares de intensidad (x_1, x_2) en el Plano (R, P)

Los pares de intensidades (x_1, x_2) se trazan mediante las Ecuaciones (30) y (33) como pares de números (R, P) . Este trazado, que proporciona una representación ponderada por la significación de los factores de cambio, se ilustra en la Fig. 3. Por simplicidad, se eligió $\sigma_1 = \sigma_2 = \sigma$.

La Fig. 3 muestra las líneas para x_1 constante y x_2 constante en el plano $(\log(R), P)$. La figura es simétrica por debajo del intercambio $R \rightarrow (1/R)$ (esto es, $\log(R) \rightarrow -\log(R)$), porque se eligió el valor $\sigma_1 = \sigma_2$ para este ejemplo en particular. Para un R dado, el intervalo de P es finito, con una cota superior $P_u(R)$ tal que:

$$0 \leq P \leq P_u(\hat{R}). \quad (34)$$

Debido a que $P_u(R) \rightarrow 0$ a medida que $R \rightarrow \infty$ ó $R \rightarrow 0$, los factores de cambio grandes están necesariamente correlacionados con valores P pequeños, aunque siempre en un intervalo no cero finito.

Una expresión para $P_u(R)$ viene dada por: (véase la Ecuación (65) en el Apéndice A para la derivación del término $P_u(R)$):

$$P_u(\hat{R}) = \begin{cases} \operatorname{erfc} \left(t_m \frac{(\hat{R}^2 + \sigma_2^2 / \sigma_1^2)^{1/2}}{(1 + \sigma_2^2 / \sigma_1^2)^{1/2}} \right), & \hat{R} \geq 1, \\ \operatorname{erfc} \left(t_m \frac{(1 / \hat{R}^2 + \sigma_1^2 / \sigma_2^2)^{1/2}}{(1 + \sigma_1^2 / \sigma_2^2)^{1/2}} \right), & \hat{R} \leq 1, \end{cases} \quad (35)$$

donde erfc es la función de error complementaria (véase Abramowitz, M. y Stegun, I. A., p. 297 *Handbook of Mathematical Functions* (Dover, New York, 1972)), y con $t_m \approx 0,477$. La Ecuación (35) es estrictamente sólo asintóticamente válida para $R \ll 1$ ó $R \gg 1$, pero de hecho proporciona una excelente aproximación para todos los valores de R, como puede verse en la Fig. 3.

Para $R > 1$, la acotación $P = P_u(R)$ corresponde a la línea $x_1 = 0$ en el plano (x_1, x_2) (esto es, el eje x_2). Los puntos sobre esta acotación son los puntos para los que un factor de cambio dado R tiene la menor significación (esto es, tiene el valor de P más grande). Las líneas para R constante corresponden a arcos dentro del plano (x_1, x_2) , que se originan todos en el eje x_2 (para $R > 1$), en cuyo punto P es un máximo, y que seguidamente son asintóticos a la línea $x_2 = R x_1$, para la que P rápidamente tiende a 0.

6. Implementación - El Algoritmo PFOLD

El esquema de estimación descrito anteriormente en la presente memoria se ha implementado en un programa C++ denominado PFOLD. Para una serie dada de parámetros de entrada ($x_1, x_2, \sigma_1, \sigma_2$), especificando las dos intensidades y las desviaciones estándar correspondientes de los términos de ruido, PFOLD evalúa primero numéricamente la función de distribución $f\eta(R)$ (véase la Ecuación (16)) sobre un intervalo finito $R_{\min} \leq R \leq R_{\max}$ en los puntos sobre una malla regular $R_i = R_{\min} + i\Delta R$, $i = 0, 1, \dots, N$, donde $R_{\min}$, $R_{\max}$ y ΔR se eligen automáticamente para capturar toda la variación de la función (Fig. 1). La función de distribución acumulativa $F(R)$ (véase la Ecuación (29)) se encuentra entonces por integración numérica de $f_R(R)$, tras lo cual todos los estimadores de la Sección 5 (esto es, el factor de cambio R , los límites de confianza (R_p, R_{1-p}), y el valor P pueden evaluarse fácilmente resolviendo numéricamente las Ecuaciones (30), (31), (32) y (33), respectivamente. Para encontrar las raíces de estas ecuaciones anteriormente mencionadas, se usó un método simple de bisección. Véase, p.ej., Press, W., et al., *Numerical Recipes in C*. 2ª Edición, p. 353 (Cambridge University Press, Cambridge, 1997).

7. Simulaciones de Monte Carlo

Un problema central en el análisis de los datos de expresión es separar el factor de cambio significativo del no significativo. Para evaluar la utilidad de los estadísticos (R, P) en este procedimiento, se realizaron una serie de simulaciones de Monte Carlo (véase, p.ej., Cowan, G., *Statistical Data Analysis*, p. 41 (Claredon Press, Oxford, 1998)) que tenían como propósito aproximarse a un experimento real. Los valores de concentración n se generaron según una distribución log-normal (véase, p.ej., Cowan, G., *Statistical Data Analysis*, p. 34 (Claredon Press, Oxford, 1998)), calculando:

$$n = \exp(y), \quad (36)$$

donde y es una variable aleatoria gaussiana, generada con los parámetros:

$$\langle y \rangle = 7,25, \quad (37)$$

$$\sigma_y = 1,22, \quad (38)$$

donde $\langle y \rangle$ y σ_y , son la media y la desviación estándar de y , respectivamente. Los parámetros de las Ecuaciones (37) y (38) dan como resultado una distribución con el percentil 25, la mediana y el percentil 75 con valores:

$$n_{25} = 618, \quad (39)$$

$$n_{50} = 1408, \quad (40)$$

$$n_{75} = 3208, \quad (41)$$

respectivamente. La elección de una distribución log-normal para la concentración η viene dictada por la observación empírica de que en experimentos reales, la distribución de las intensidades de los genes presentes es aproximadamente log-normal. Para los chips Affymetrix®, con un procedimiento de tinción por anticuerpos después de la hibridación, las intensidades indicadas por los percentiles en las Ecuaciones (39), (40) y (41) son típicas.

Para cada valor de η generado por la Ecuación (36), se simuló un factor de cambio real b , combinado con el ruido, calculando los dos valores de intensidad:

$$x_1 = n + \varepsilon_1, \quad (42)$$

$$x_2 = bn + \varepsilon_2, \quad (43)$$

donde los términos de ruido ε_1 y ε_2 , son variables aleatorias gaussianas no correlacionadas con medias $\langle \varepsilon_1 \rangle = \langle \varepsilon_2 \rangle = 0$, y con desviaciones estándar dadas por la Ecuación (6) con los parámetros:

$$\sigma_c = 600, \quad \alpha = 0,25. \quad (44)$$

Finalmente, a partir de las intensidades (x_1, x_2) calculadas con las Ecuaciones (42) y (43), se calcularon los estimadores correspondientes (R, F) usando las Ecuaciones (30) y (33).

7.1 Asignaciones de Clase

Para determinar el grado de seguridad, se llevaron a cabo dos series de simulaciones, cada una de las cuales definía una clase de genes sufriendo un factor de cambio dado:

Clase 0: sin cambio, $b = 1$.

Clase 1: cambio, $b = 3$.

La eficacia de usar PFOLD para distinguir entre las 2 clases de genes se evaluó para seleccionar los genes que pertenecen a la Clase 1. Para realizar esta evaluación, se definió una región de aceptación D dentro del plano (R, F) (véase, p.ej., Cowan, G., *Statistical Data Analysis*, p. 47 (Claredon Press, Oxford, 1998)), así como la definición de la predicción π para la pertenencia a una clase de un gen como:

$$\pi = \begin{cases} p \text{ si } (\hat{R}, P) \in D_1, & \text{(esto es, asignar gen a la clase 1)} \\ \alpha \text{ si } (\hat{R}, P) \in D_0 & \text{(esto es, asignar gen a la clase 0)} \end{cases} \quad (45)$$

donde p y α significan la presencia y ausencia en la región de aceptación, respectivamente. Un ejemplo de una región de aceptación D es una con una superficie de decisión rectangular definida por:

$$D = \{ \hat{R} \geq R_c, P \leq P_c \}, \quad (46)$$

sin embargo, se considerarán también regiones más generales.

Para cualquier elección de D , es posible derivar un valor estimado de las probabilidades:

$P(p | 0)$ = probabilidad de que un gen en la Clase 0 se asigne a la Clase 1,

$P(\alpha | 0)$ = probabilidad de que un gen en la Clase 1 se asigne a la Clase 0,

simplemente contando, en cada caso, el número de casos (R, P) que pertenecen o, alternativamente, no pertenecen a D . Si a su vez, los valores se suponen para las probabilidades *a priori* como que, en una serie grande de genes, algunos no sufrirán un factor de cambio, y otros un factor de cambio de 3 (por simplicidad, se supone que no son posibles otros factores de cambio fuera de 1 ó 3),

P_0 = probabilidad *a priori* de que un gen esté en la clase 0,

P_1 = probabilidad *a priori* de que un gen esté en la clase 1,

entonces puede utilizarse el Teorema de Bayes para obtener las probabilidades *a posteriori*:

$P(p | 0)$ = probabilidad de que un gen asignado a un factor de cambio significativo no haya cambiado realmente,

$P(\alpha | 0)$ = probabilidad de que un gen asignado a la categoría sin cambio realmente cambió.

El resultado es:

$$P(0 | p) = P_0 P(p | 0) / P_p, \quad (47)$$

$$P(1 | \alpha) = P_1 P(\alpha | 1) / P_u. \quad (48)$$

donde P_p y P_u , las probabilidades *a posteriori* totales de declarar un gen en la Clase 1 ó Clase 0, respectivamente, vienen dadas por:

$$P_p = P(p | 0)P_0 + (1 - P(\alpha | 1))P_1, \quad (49)$$

$$P_u = (1 - P(p | 0))P_0 + P(\alpha | 1)P_1. \quad (50)$$

Las dos cantidades que son de mayor interés incluyen:

$$P(\alpha | 1) = \text{ratio absoluto de falsos negativos}, \quad (51)$$

$$P(0 | p) = \text{ratio relativo de falsos positivos}. \quad (52)$$

Las definiciones de las Ecuaciones (51) y (52) no son simétricas porque el cálculo de $P(0 | p)$ requiere el valor del P_1 anterior, mientras que el cálculo de $P(\alpha | 1)$ no. El ratio absoluto de falsos negativos es una medida de la fracción de todos los genes expresados que no se tendrán en cuenta por el esquema de detección en unas condiciones de rigurosidad dadas. El ratio relativo de falsos positivos, por otro lado, se refiere a la fracción de los genes detectados que se clasifica erróneamente, y que no sufrió de hecho un cambio. Por tanto, el ratio de falsos negativos, como se define en la presente memoria, es así una medida de la eficiencia (véase, p.ej., Cowan, G., *Statistical Data Analysis*, p. 47 (Claredon Press, Oxford, 1998)), o sensibilidad del esquema de detección (esto es, se prefiere el valor más pequeño), mientras que el ratio de falsos positivos es una medida de la pureza (véase, p.ej., Cowan, G., *Statistical Data Analysis*, p. 47 (Claredon Press, Oxford, 1998)) o selectividad del esquema (esto es, se prefiere el valor más pequeño).

7.2 Resultados de la simulación experimental

En la Fig. 6 (Paneles A y B), hay una comparación de gráficos de dispersión dentro del plano (R, P) generado por 1000 genes de la Clase 0 (la clase sin cambio) y 1000 genes en la Clase 1 (los genes que mostraron un factor de cambio 3), respectivamente. Para facilitar la comprensión de este concepto, las siguientes ecuaciones se basan en un escenario con una probabilidad *a priori* de que un gen cambie de:

$$P_1 = 0,2. \quad (53)$$

Esto define una configuración con 1000 genes cambiando con un factor de 3, frente a un fondo de 4000 genes no cambiantes.

Un método de elección para mostrar la dependencia de las tasas de error en la posición de una superficie de decisión es la construcción, en formato gráfico, de la denominada característica del operador del receptor (ROC) del esquema de decisión. Véase, p.ej., Van Trees, H. L., *Detection, Estimation and Modulation Theory*, Part I (John Wiley and Sons), New York, 1998). La ROC permite visualizar claramente el equilibrio entre reducir, p.ej., minimizar, ratios de falsos positivos, aumentando así la selectividad del esquema, y reducir, p.ej., minimizar, los ratios de falsos negativos, aumentando así también la sensibilidad global del esquema dado.

En la Fig. 7, se representan las características del operador del receptor (ROC) que se obtienen cuando se usa como estadístico para una acotación de decisión el mismo factor de cambio, $t = R$. Por lo tanto, en el caso anterior, la región de aceptación se define simplemente como:

$$D = \{\hat{R} \geq R_c\}. \quad (54)$$

de manera que la superficie de decisión es una línea vertical dentro del plano (R, P). En la Fig. 7, el ratio de falsos positivos $P(0 | p)$ y el ratio de falsos negativos $P(\alpha | 1)$ se representan como una función de R_c . Puede observarse que a medida que se aumenta la rigurosidad del test para la aceptación (esto es, R_c creciente), el ratio relativo de falsos positivos disminuye (esto es, la pureza de la muestra detectada aumenta), pero que hay asimismo un aumento concomitante en el ratio absoluto de falsos negativos también. Además, la Fig. 8 ilustra la característica del operador del receptor (ROC) que se obtiene para el estadístico $t = -P$.

La Fig. 5 ilustra en forma tabular la realización de las tres metodologías estadísticas utilizadas para la detección de 1000 genes con un factor de cambio real $b = 3$, frente a un fondo de 4000 genes sin factor de cambio $b = 1$, basadas en las simulaciones experimentales de Monte Carlo descritas en la Sección 7 (esto es, la probabilidad *a priori* de que un gen sufra un factor de cambio 3 es $P_1 = 0,2$). Se presentan resultados específicos para un ratio de falsos positivos fijo $P(0 | p) = 0,3$. La Fig. 9 ilustra la sensibilidad comparada como una función de P_1 = fracción de genes que sufrieron un cambio ($b = 3$), para los estadísticos R y P. Se impone un ratio relativo de falsos positivos fijo $P(0 | p) = 0,3$ de principio a fin.

Los resultados ilustrados en la Fig. 5 y la Fig. 7 muestran que el uso del valor P (t = estadístico -t) en lugar de, o en conjunción con, el factor de cambio R, puede aumentar notablemente la sensibilidad de los diversos intervalos de parámetros que eran, hasta ahora, extremadamente difíciles de detectar.

En conclusión, debido a la baja relación de señal frente a ruido que es inherente a los datos de expresión, este ruido debe tenerse en cuenta cuidadosamente. El algoritmo PFOLD descrito en la presente memoria ofrece un entorno teórico y práctico para tratar el ruido. Por ejemplo, el algoritmo PFOLD tiene en cuenta dos medidores importantes para el factor de cambio del nivel de expresión génica: (i) el valor P, que refleja la “calidad” global de la relación; y (ii) R, que refleja la “cantidad” del factor de cambio en la expresión del(los) gen(es). Además, el estadístico p de PFOLD es esencial para cuantificar pequeñas poblaciones de genes cambiantes y/o pequeños factores de cambio del nivel de expresión de tales genes.

Apéndice A: Distribución para $X_1 = 0$

Para un factor de cambio dado $R > 1$, las predicciones menos significativas surgen de un par de intensidades con $x_1 = 0$. Para investigar esta dependencia, se genera la forma aproximada de la Ecuación (16) cuando $x_1 = 0$, y $x_2 \gg \sigma_{1,2}$. El resultado es:

$$f_R(R) \approx \left(\frac{2}{\pi}\right)^{1/2} \frac{Ry}{(R^2 + \sigma_2^2 / \sigma_1^2)^{3/2}} \exp\left(-\frac{y^2}{2(R^2 + \sigma_2^2 / \sigma_1^2)}\right), \quad (55)$$

donde $y = x_2 / \sigma_1$. Con la transformación:

$$u = \frac{1}{(R^2 + \sigma_2^2 / \sigma_1^2)^{1/2}}, \quad (56)$$

el intervalo $0 \leq R < \infty$ se traza en $0 < u < \sigma_1 / \sigma_2$, y la distribución para u viene dada por:

$$f_u(u) = \left(\frac{2}{\pi}\right)^{1/2} y \exp(-y^2 u^2 / 2). \quad (57)$$

Es fácil comprobar que se obtiene la correcta normalización de la función de distribución, puesto que:

$$\int_0^\infty f_R(R) dR = \int_0^{\sigma_1 / \sigma_2} f_u(u) du \approx \int_0^\infty f_u(u) du = 1. \quad (58)$$

La aproximación en la Ecuación (58) es válida porque $y = x_2 / \sigma_1 \gg 1$.

Para calcular el valor P para un valor dado de y:

$$P = P(R \leq 1) = P(u \geq u_1), \quad (59)$$

donde:

$$u_1 = (1 + \sigma_2^2 / \sigma_1^2)^{1/2}. \quad (60)$$

Usando la Ecuación (57), e integrando sobre el intervalo $u_1 \leq u < \infty$, se encuentra que:

$$P = \operatorname{erfc} \left(\frac{y}{2^{1/2}} (1 + \sigma_2^2 / \sigma_1^2)^{-1/2} \right). \quad (61)$$

Para calcular el factor de cambio predicho R para un valor dado de y, se utiliza la definición de R como la mediana de la distribución:

$$\frac{1}{2} = P(R \leq \hat{R}) = P(u \geq). \quad (62)$$

Esto da como resultado la ecuación:

$$\frac{1}{2} = \text{erf} \left(\frac{y}{2^{1/2}} \left(\hat{R}^2 + \sigma_2^2 / \sigma_1^2 \right)^{-1/2} \right), \quad (63)$$

a partir de la cual encontramos la relación entre y y R:

$$y = 2^{1/2} t_M \left(\hat{R}^2 + \sigma_2^2 / \sigma_1^2 \right)^{1/2}, \quad (64)$$

donde $t_M \approx 0,477$ es la raíz de la ecuación $\text{erf}(t_M)=1/2$. Usando la Ecuación (64), y puede eliminarse de la Ecuación (61), obteniéndose la ecuación final:

$$P_u(\hat{R}) = \begin{cases} \text{erfc} \left(t_M \frac{\left(\hat{R}^2 + \sigma_2^2 / \sigma_1^2 \right)^{1/2}}{\left(1 + \sigma_2^2 / \sigma_1^2 \right)^{1/2}} \right), & \hat{R} \geq 1, \\ \text{erfc} \left(t_M \frac{\left(1 / \hat{R}^2 + \sigma_1^2 / \sigma_2^2 \right)^{1/2}}{\left(1 + \sigma_1^2 / \sigma_2^2 \right)^{1/2}} \right), & \hat{R} < 1, \end{cases} \quad (65)$$

donde $t_M \approx 0,477$.

Con referencia a la Fig. 11, se muestra un procedimiento 150 para calcular diferencias en el nivel de expresión génica en al menos una matriz de genes. El procedimiento 150 usa una fórmula matemática para describir una distribución estadística *a posteriori* de los niveles (p.ej., todos los niveles) de expresión génica que pueden derivarse de las mediciones obtenidas de los niveles de expresión génica en uno o más tipos de células o tejidos representados en al menos una matriz.

En la etapa 152, se define un ruido experimental asociado con la intensidad de la señal de hibridación para cada gen en al menos una matriz. Este ruido es ruido experimental, siendo las variaciones en los niveles observados en los chips u otras micromatrices más que el ruido biológico, que es la variación del nivel de expresión visto en sistemas biológicos.

En la etapa 154, el ruido experimental definido se usa para definir una función de distribución de probabilidad (pdf) analítica que describe los valores de distribución de la intensidad para cada gen. Se supone que el ruido es gaussiano y la pdf analítica se define usando el Teorema de Bayes. La pdf analítica es una función continua.

En la etapa 156, las pdfs analíticas se usan para derivar una pdf conjunta analítica que describe las posibles relaciones o factores de cambio para genes o productos génicos expresados de manera diferencial en al menos una matriz. La pdf conjunta analítica que describe las posibles relaciones o factores de cambio puede derivarse para cualesquier genes o productos génicos expresados de manera diferencial.

En la etapa 158, las pdfs conjuntas se aplican, usando las intensidades y valores de ruido derivados experimentalmente de los genes en al menos una matriz, para determinar los valores relacionados con los factores de cambio asociados con los genes. Se estiman los factores de cambio en la concentración de los productos de transcripción génica. Se establecen los límites de confianza sobre el factor de cambio, dados unos intervalos de confianza específicos. Se deriva también un valor p, o medidor de la calidad asociado con el valor estimado del factor de cambio. Este valor representa la probabilidad de que un factor de cambio pudiera ser menor que 1 cuando el valor estimado es mayor que 1, o la probabilidad de que el factor de cambio sea mayor que 1 cuando el valor estimado es menor que 1.

El factor de cambio estimado representa la diferencia en el nivel de expresión génica observado. La varianza total (ruido) puede aún ser alta incluso a medida que la concentración del producto de transcripción se aproxima a cero.

REIVINDICACIONES

1. Un método para calcular indicios de diferencias en el nivel de expresión génica en una pluralidad de hibridaciones de matrices, comprendiendo el método:
 - (a) determinar una indicación del ruido experimental asociado con la intensidad de la señal de hibridación para un gen en cada hibridación de matrices;
 - (b) usar la indicación del ruido experimental determinada para determinar una función de distribución de probabilidad analítica que describe los valores de distribución de la intensidad para el gen en cada hibridación de matrices aplicando el Teorema de Bayes;
 - (c) usar las funciones de distribución de probabilidad analíticas para derivar una función de distribución de probabilidad conjunta analítica que describe relaciones posibles o factores de cambio para un gen expresado de manera diferencial en dicha pluralidad de hibridaciones de matrices; y
 - (d) aplicar la función de distribución de probabilidad conjunta del gen expresado de manera diferencial usando las intensidades y valores de ruido derivados experimentalmente de las hibridaciones de matrices, para determinar los valores relacionados con el factor de cambio en el nivel de expresión del gen expresado de manera diferencial.
2. El método de la reivindicación 1, en el que los valores determinados incluyen los factores de cambio estimados en la concentración de los productos de transcripción génica.
3. El método de la reivindicación 2, en el que los valores determinados incluyen un medidor de la calidad asociado con al menos uno de los factores de cambio estimados.
4. El método de la reivindicación 3, en el que el medidor de la calidad representa al menos una de una probabilidad de que un factor de cambio pudiera ser menor que 1 cuando un factor de cambio estimado es mayor que 1, y una probabilidad de que un factor de cambio sea mayor que 1 cuando el factor de cambio estimado es menor que 1.
5. El método de la reivindicación 1, en el que los valores determinados incluyen los límites de confianza sobre el factor de cambio dados intervalos de confianza específicos.
6. El método de la reivindicación 1, en el que la función de distribución de probabilidad analítica se define suponiendo que el ruido es gaussiano.
7. Un producto programa de ordenador, para calcular indicios de diferencias en el nivel de expresión génica en una pluralidad de hibridaciones de matrices, que reside en un medio legible en ordenador, y que comprende instrucciones que hacen al ordenador:
 - (a) determinar una indicación del ruido experimental asociado con la intensidad de la señal de hibridación para un gen en cada hibridación de matrices;
 - (b) usar la indicación del ruido experimental determinada para determinar una función de distribución de probabilidad analítica que describe los valores de distribución de la intensidad para el gen cada hibridación de matrices aplicando el Teorema de Bayes;
 - (c) usar las funciones de distribución de probabilidad analíticas para derivar una función de distribución de probabilidad conjunta analítica que describe relaciones posibles o factores de cambio para un gen expresado de manera diferencial en dicha pluralidad de hibridaciones de matrices; y
 - (d) aplicar la función de distribución de probabilidad conjunta del gen expresado de manera diferencial usando las intensidades y valores de ruido derivados experimentalmente de las hibridaciones de matrices, para determinar los valores relacionados con el factor de cambio en el nivel de expresión del gen expresado de manera diferencial.
8. El producto programa de ordenador de la reivindicación 7, en el que los valores determinados incluyen los factores de cambio estimados en la concentración de los productos de transcripción génica.
9. El producto programa de ordenador de la reivindicación 8, en el que los valores determinados incluyen un medidor de la calidad asociado con al menos uno de los factores de cambio estimados.
10. El producto programa de ordenador de la reivindicación 9, en el que el medidor de la calidad representa al menos una de una probabilidad de que un factor de cambio pudiera ser menor que 1 cuando un factor de cambio estimado es mayor que 1, y una probabilidad de que un factor de cambio sea mayor que 1 cuando el factor de cambio estimado es menor que 1.

ES 2 331 737 T3

11. El producto programa de ordenador de la reivindicación 7, en el que los valores determinados incluyen los límites de confianza sobre el factor de cambio dados intervalos de confianza específicos.

12. El método de la reivindicación 1, en el que se determina una indicación del ruido experimental, se determina la función de distribución de probabilidad, y se deriva la función de distribución de probabilidad conjunta y se aplica para cada gen en cada hibridación de matrices.

13. El producto programa de ordenador de la reivindicación 7, en el que las instrucciones para hacer al ordenador determinar una indicación de ruido experimental, determinar la función de distribución de probabilidad, derivar la función de distribución de probabilidad conjunta, y aplicar la función de distribución de probabilidad conjunta, hacen que el ordenador lo haga para cada gen en cada hibridación de matrices.

14. El producto programa de ordenador de la reivindicación 7, en el que la función de distribución de probabilidad analítica se determina suponiendo que el ruido es gaussiano.

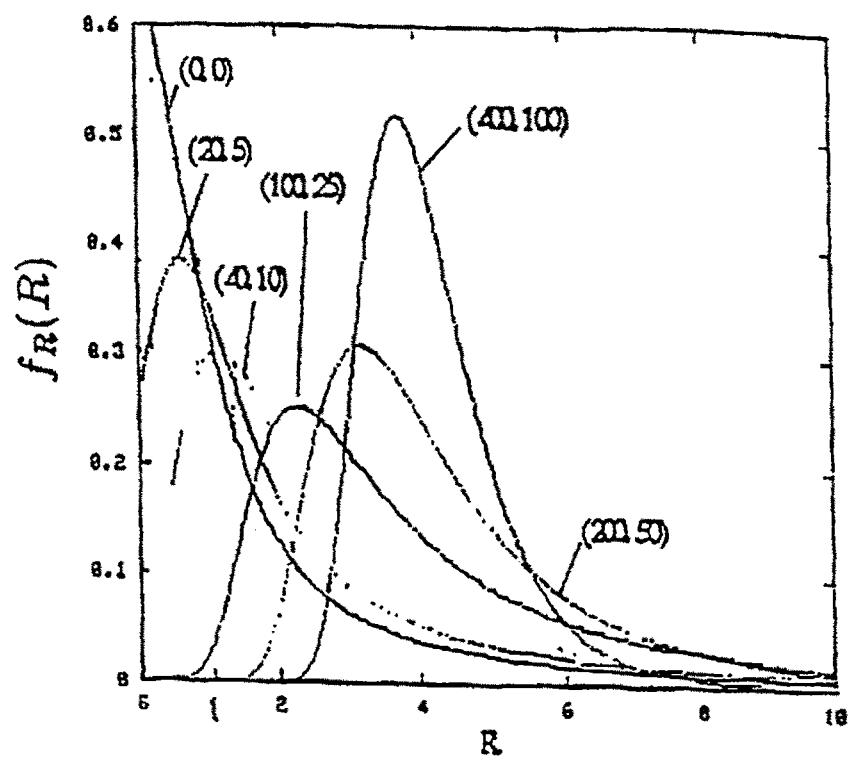


FIG. 1

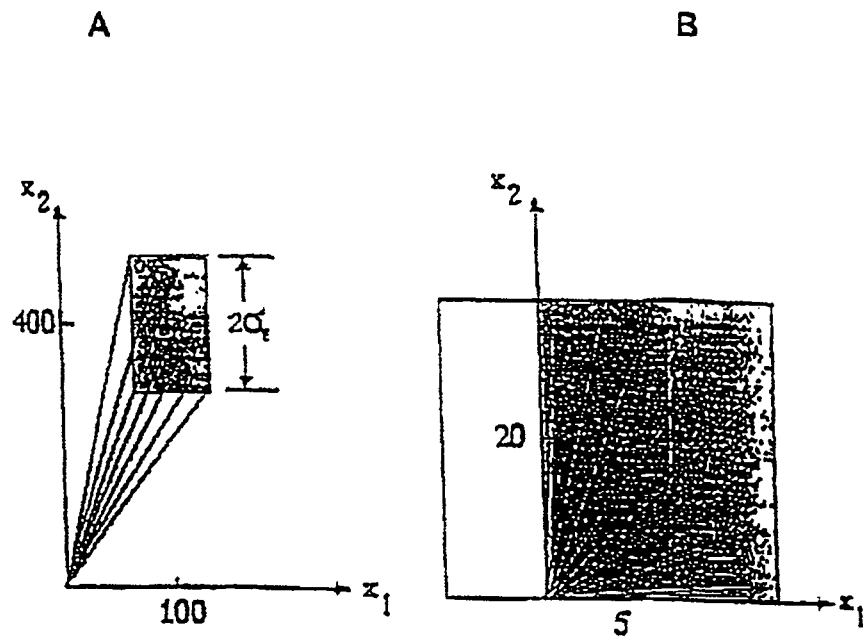


FIG. 2

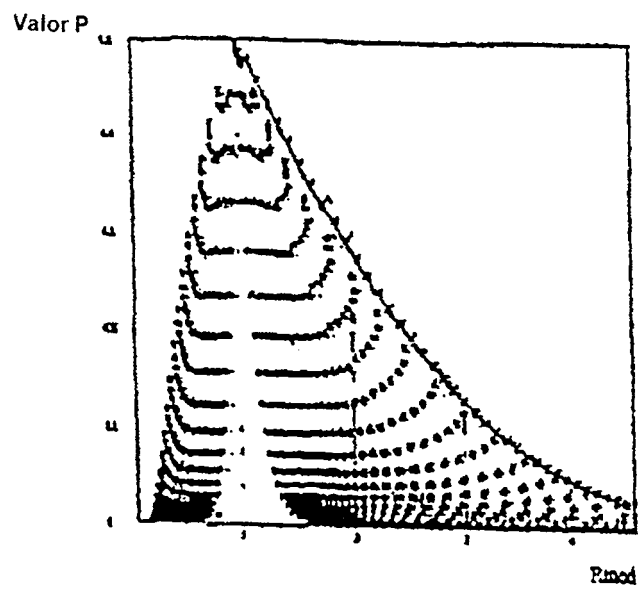


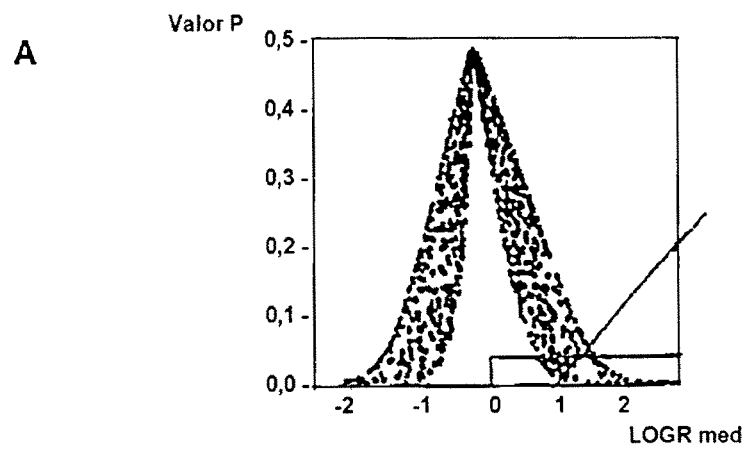
FIG. 3

X_1	X_2	R_P	R	R_{1-P}	Valor P
100	400	3,3	4,0	5,0	0,0
50	200	2,8	3,9	6,5	5,98X10
25	100	2,0	3,6	8,8	4,4JX10
10	40	0,93	2,2	7,3	0,18
5	20	0,51	1,5	5,-	0,34
1	4	0,32	1,1	4,6	0,46
0	0	0,23	1,0	4,4	0,5

Fig. 4

Parámetros de Decisión		Resultados para ratio de F.P. constante $P(0/p) = 0,3$				
Estadístico principal t	Región	N_{TOT}	NTP	Ratio T.P. $P(p/t)$	t_c	Med (R)
Factor de cambio R	$R \geq t_c$	371	260	0,26	4,37	6,1
Valor P	$P \leq t_c$ $R > 1$	814	570	0,57	0,037	4,2
Discriminante $\log(R) - 5,48P$	$t \geq t_c$	713	499	0,50	0,95	4,4

Fig. 5



Gráfica de dispersión

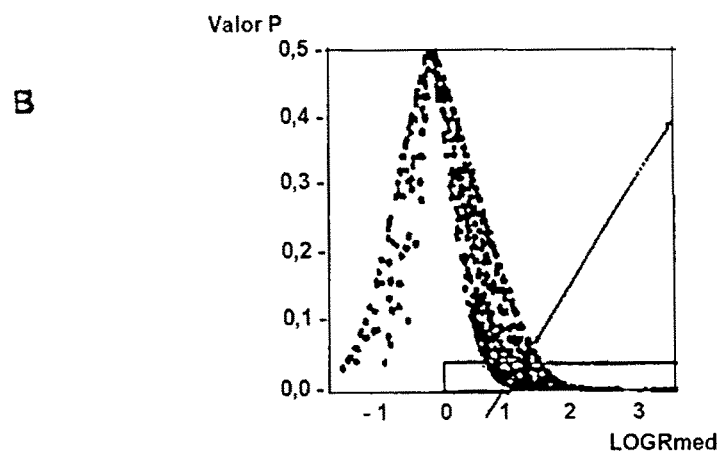


FIG. 6

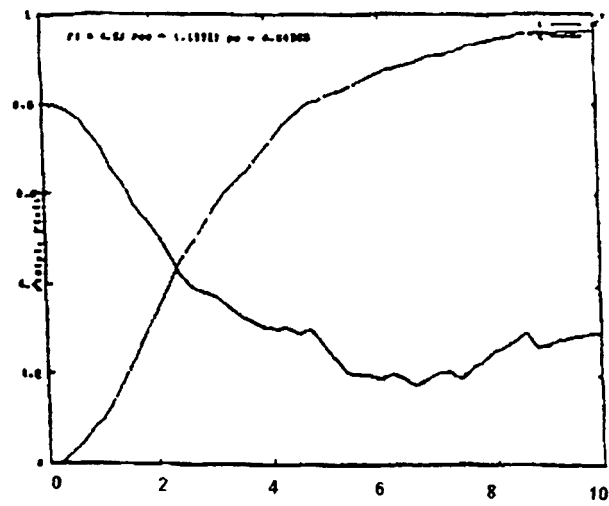


FIG. 7

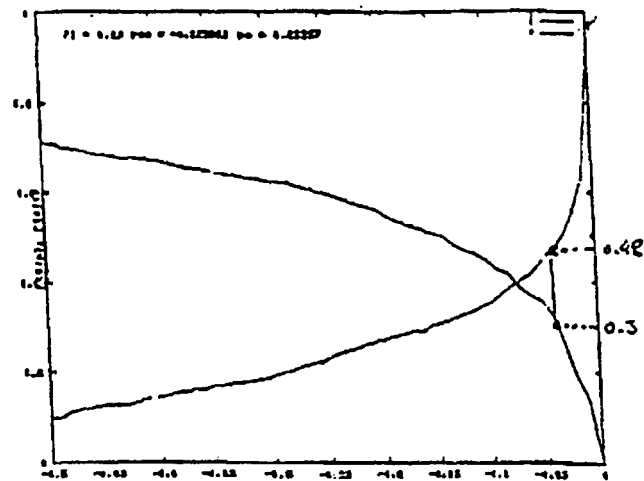


FIG. 8

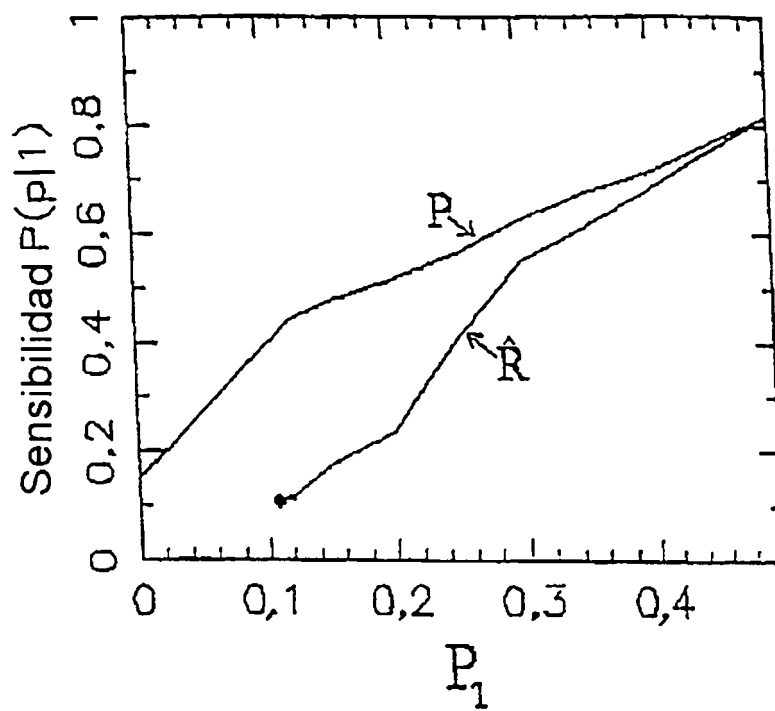


FIG. 9

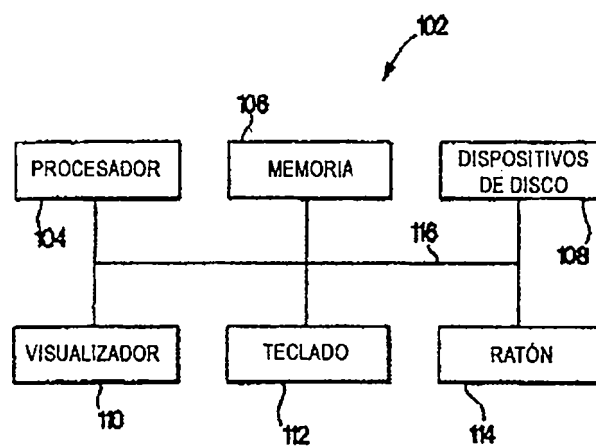


Fig. 10

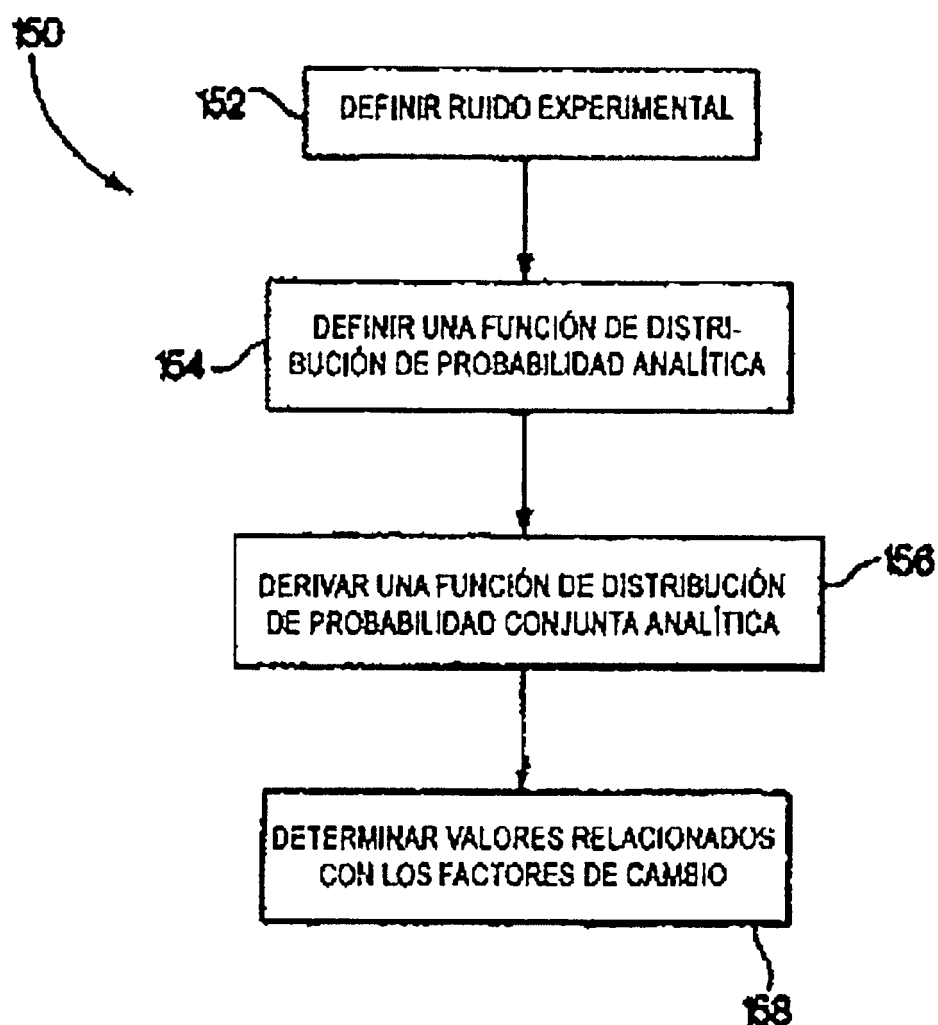


Fig. 11