Microarray: le matrici a sonde

O. Valentini, B. Milanesi

Laboratori di Patologia Clinica, Azienda Ospedaliera di Desenzano del Garda (BS)

Riassunto

I "microarray" costituiscono lo strumento più sofisticato e versatile per l'analisi quantitativa di geni e prodotti genici. Con una sola analisi è possibile raccogliere una quantità impressionante di dati, che può essere archiviata in forma digitale e confrontata con dati precedenti o successivi per poter dedurre, per esempio, quali geni sono coinvolti, e in quale misura, in determinati processi fisiologici, patologici o indotti da farmaci. Le applicazioni cliniche sono molteplici, in particolare nell'ambito della diagnostica delle malattie neoplastiche, infettive, allergiche o autoimmuni. Le informazioni ottenibili sul singolo paziente consentono di definire in dettaglio il quadro diagnostico e prognostico, nonché di stabilire quale sia la terapia farmacologica più efficace e meno tossica.

Summary

Microarrays: an overview

Microarrays provide an unprecedented opportunity for comprehensive concurrent analysis of thousands of genes and gene products. Through a single analysis it is possible to collect a huge mass of data that can be stored in digital form and compared with previous and successive data in order to deduce, for example, which genes, and to what extent, are involved in physiological, pathological or drug-induced processes. They have many and various clinical applications, particularly in the diagnostics of neoplastic, infectious, allergic or autoimmune diseases. Data obtainable for a single patient make it possible to define in detail the diagnostic and prognostic picture and to choose the least toxic and most effective therapy.

Premessa

La tumultuosa evoluzione tecnologica degli ultimi anni sia in biologia, con l'analisi di sequenza dell'intero genoma umano, sia nella fisica informatica, con la realizzazione di elaboratori elettronici capaci di miliardi di operazioni al secondo, grazie all'estrema miniaturizzazione dei circuiti a semiconduttore, ha creato le premesse per una interazione tra discipline diverse, come appunto la fisica, l'ingegneria, l'informatica, la matematica e la biologia, che ha portato allo sviluppo dei "microarray".

Il termine "microarray" sta per "microscopic glass array", ossia disposizione ordinata, o schieramento, su un vetrino da microscopio, di elementi, o sonde, che consentono il legame specifico di geni o prodotti genici.

Che cos'è un "microarray"? In estrema sintesi e in maniera non esaustiva lo si può definire un dispositivo per l'analisi biologica formato da un sistema di sonde in grado di legare in modo specifico acidi nucleici oppure proteine per ottenere attraverso un singolo passaggio analitico una enorme mole di informazioni estremamente dettagliate e tali da costituire una sorta di "fotografia" dello stato funzionale della cellula nel momento dell'analisi. Le singole sonde di cui il dispositivo è costituito sono legate al supporto su piccole aree circolari disposte secondo un preciso ordine geometrico lungo linee orizzontali (righe) e linee verticali (colonne), in modo del tutto analogo a quello dei numeri di una matrice (in matematica si definisce "matrice" un insieme di numeri ordinati in una tabella rettangolare in modo da formare *m* righe e *n* colonne). Ogni singola

Ricevuto: 16-04-2007 Accettato: 29-10-2007 Pubblicato on-line: 23-11-2007

Corrispondenza a: Dott. Omar Valentini, Laboratorio di Patologia Clinica, Ospedale Civile di Gavardo, Dipartimento di Medicina di Laboratorio, Azienda Ospedaliera di Desenzano del Garda, Via Andrea Gosa n. 74, 25085 Gavardo (BS).
Tel. 0365-378420, fax 0365-378235, e-mail: omar.valentini@aod.it

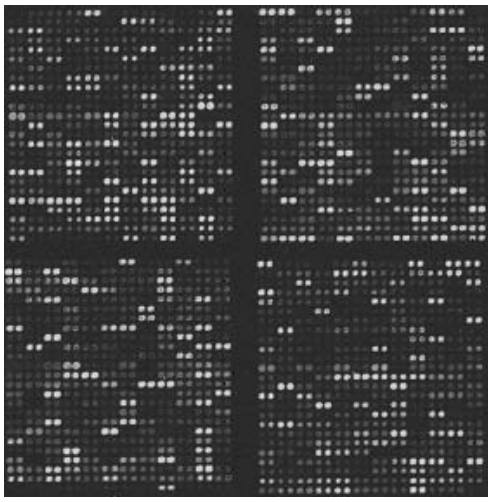

Figura 1. Immagine archiviata in forma digitale di una matrice a DNA. Gli elementi della matrice risultano più o meno luminosi in relazione alla quantità della specifica molecola bersaglio legata.

area circolare rappresenta un elemento della matrice. Ad ogni elemento corrisponde una singola specie di sonda. Gli elementi della matrice contengono tutti una uguale quantità molare di sonda e sono distribuiti sul supporto in modo da restare equidistanti lungo entrambe le direttrici ortogonali (Fig. 1). Ogni elemento è individuabile in modo univoco attraverso le sue coordinate (ascissa e ordinata). Mentre gli elementi di una matrice matematica sono rappresentati da numeri, quelli di un "microarray" sono costituiti da sonde specifiche e per questo motivo può essere giustificato il termine di "matrici a sonde". Anche il termine alternativo di "sonde a schiera" appare adeguato: schiera è una moltitudine ordinata in lunghe file una accanto all'altra o comunque raggruppata secondo determinati o prestabiliti criteri. Da qualcuno viene usato anche il termine di "griglia", di "chip" o di "biochip".

Ogni elemento della matrice, corrispondente a una singola specie molecolare di identità nota (acido nucleico, proteina o tessuto, a seconda del tipo di "microarray"), legandosi a molecole (acidi nucleici o proteine) presenti nel campione saggiato è in grado di fornire informazioni di tipo qualitativo e, in determinate condizioni, anche di tipo quantitativo, in numerosi ambiti della ricerca biologica. I "microarray" consentono di passare dalla fase di conoscenza dell'"anatomia" del genoma, rappresentata dalla sequenza del DNA, a quella della "fisiologia" del genoma stesso. Attraverso di essi è possibile ottenere dati di "genomica funzionale", ovvero dati relativi all'attività dei singoli geni di una cellula. Una fibrocellula muscolare e una cellula nervosa di uno stesso organismo hanno lo stesso corredo genetico, ma fenotipo evidentemente assai diverso. La differenziazione cellulare è il risultato di un processo di attivazione e disattivazione selettiva di geni, dove per attività di un gene si intende la quantità di RNA messaggero trascritto a partire dalla specifica sequenza di DNA. I "microarray" a DNA, o matrici a DNA, legando gli mRNA della cellula, ne misurano, in opportune condizioni sperimentali, la quantità relativa e forniscono così una immagine dell'attività dei suoi geni, cioè dell'attività trascrizionale della cellula stessa. Ciò che si ottiene è una sorta di "foto istantanea" del quadro di espressione genica.

Se si considerano nel loro insieme i processi attraverso cui viene modulata la sintesi proteica, nel passaggio da corredo genetico, o genoma, a corredo proteico, o proteoma, non si può sottacere il fatto che l'analisi con "microarray" a DNA fornisce informazioni solo su un primo stadio della modulazione, cioè quello che riguarda il processo di trascrizione dallo stampo di DNA all'RNA messaggero complementare, mentre non ne fornisce sul processo di traduzione. Questo secondo stadio della modulazione, che interessa il passaggio dall'RNA messaggero alla catena proteica, è di importanza minore, ma non sempre trascurabile (il problema, come si potrà in seguito intuire, è parzialmente risolto dall'utilizzo di specifiche matrici a proteine). Nondimeno i dati ottenibili dall'analisi condotta a livello trascrizionale sono di enorme importanza in ambito biologico e medico e non potrebbero essere raccolti in maniera altrettanto efficace con altri mezzi. Le matrici, opportunamente progettate e realizzate, consentono infatti di valutare l'attività di tutti i geni di una cellula in un unico passaggio, ossia attraverso una singola analisi. L'enorme mole di dati prodotti non è sempre immediatamente utilizzabile e un congruente sistema di lettura, nonché una valutazione critica dei dati stessi sono essenziali per estrapolare da essi l'effettiva realtà biologica. Questi aspetti possono essere pienamente compresi solo entrando nel dettaglio tecnico dell'analisi e facendo riferimento agli studi critici riportati nella letteratura scientifica. Oltre che per l'analisi dell'espressione genica, si possono progettare e utilizzare matrici a DNA per l'analisi dei polimorfismi e l'analisi di sequenza del genoma.

Cenni storici

La tecnologia della "matrice", o "microarray", che in forma del tutto embrionale si può in qualche modo far risalire agli anni Settanta, è stata sviluppata appieno a partire dall'inizio degli anni Novanta all'Università di Stanford ad opera di Mark Schena e collaboratori, cui va il merito di averla fatta evolvere fino alla forma attuale. La prima pubblicazione in cui venga descritto un sistema analitico che in qualche modo precorre le attuali matrici a DNA è di Grunstein e Hogness (1975)[1], che utilizzarono schiere di DNA immobilizzato su filtro di nitrocellulosa ottenuto da colonie di *E. coli* contenenti plasmidi per la clonazione genica. Davis e coll. svilupparono anch'essi saggi su filtri di nitrocellulosa utilizzati per la selezione di cloni ricombinanti[2,3]. Studi più innovativi di ibridazione su supporto solido vennero condotti a Mosca tra la fine degli anni Ottanta e

l'inizio dei Novanta da ricercatori del gruppo di Mirzabekov, che descrissero la possibilità di sequenziare il DNA con tecnica ibridativa utilizzando matrici a oligonucleotidi legati a gel di poliacrilamide fissato su vetrino[4-6]. Fodor e coll. (1991)[7] descrissero un procedimento di sintesi chimica *in situ* di oligonucleotidi associata all'impiego di tecniche di tipo fotolitografico. Maskos e Southern idearono ancora nuove tecniche per la sintesi *in situ* su supporto di vetro di sonde oligonucleotidiche, utilizzando le matrici così ottenute per studi di ibridazione[8,9]. Lerach e coll. misero a punto i primi sistemi robotizzati per la deposizione di sonde su supporto solido, nella fattispecie di nylon; l'evoluzione tecnologica in questo ambito fu poi assai rapida[10]. L'opera così iniziata nel suo complesso fu portata a compimento da Schena e coll.[11,12], che tracciarono la via per l'utilizzo delle matrici nell'analisi dell'espressione genica, che a tutt'oggi resta il loro impiego principale. Lo sviluppo dei supporti per "microarray" è andato di pari passo con quello dei microprocessori a semiconduttore, attualmente capaci di operare miliardi di operazioni al secondo (capacità operative dell'ordine dei gigahertz) su circuiti miniaturizzati delle dimensioni di frazioni di micrometro. I primi supporti per matrice, nel 1995, contenevano 96 geni, legati ognuno a una superficie del diametro di 200 micrometri; nel 2001 contenevano già 30.000 geni, ognuno occupante una superficie del diametro di 16 micrometri. Attualmente sono disponibili "microarray" per l'intero genoma umano, che rappresentano circa 41.000 geni e trascritti e che rendono possibile lo studio completo dell'espressione genica cellulare attraverso un singolo passaggio analitico.

Come funzionano le matrici?

Il funzionamento delle matrici si basa sull'interazione di legame tra biomolecole complementari, ovvero tra le sonde fissate alla matrice e le molecole ad esse complementari (ossia in grado di formare con esse legami a elevata stereospecificità) presenti in un campione biologico. Le più importanti sonde di questo tipo sono quelle a DNA, ma è possibile utilizzare a scopo analitico anche matrici aventi come sonde proteine (comprese quelle di tipo anticorpale)[13-16], peptidi o anche specifici leganti di basso peso molecolare, come per esempio substrati di reazioni enzimatiche (che si legano in modo stereospecifico al sito catalitico di un determinato enzima), o come per esempio apteni, che possono riconoscere, legandosi ad essi, gli specifici anticorpi. Esistono peraltro matrici i cui elementi sono rappresentati da sottili fettine di tessuto[17-22]. Per semplicità verranno qui illustrati alcuni aspetti inerenti le matrici a DNA, data la loro maggiore importanza.

Nell'ambito delle matrici a DNA possiamo distinguere, in relazione alla lunghezza delle sonde utilizzate, due diverse tipologie: matrici a cDNA e matrici a oligonucleotidi (oDNA). Le prime vengono ottenute deponendo con micropipette e legando chimicamente su vetrino le singole sonde, ognuna rappresentata da uno specifico clone di cDNA (DNA retrotrascritto su uno specifico mRNA), tipicamente della lunghezza di 500-1000 (a volte fino a 2500) coppie di basi. Questo tipo di matrice, caratterizzato da una bassa densità di schieramento (100-500 geni), è stato ed è tuttora quello più ampiamente usato nei saggi di espressione genica dai quali si vogliano ottenere dati più rigorosamente quantitativi. I lavori scientifici pubblicati su matrici a cDNA rappresentano circa il 65% del totale delle pubblicazioni su "microarray". Le matrici a oligonucleotidi si preparano sintetizzando *in situ* sul supporto di vetro le sonde oligonucleotidiche (a singola elica e della lunghezza di 15-80 nucleotidi) attraverso un procedimento che associa la sintesi chimica con tecniche fotolitografiche simili a quelle utilizzate nella fabbricazione miniaturizzata dei semiconduttori (microprocessori), oppure deponendo con micropipette oligonucleotidi già sintetizzati e legandoli chimicamente alla matrice. Questo tipo di matrici è di norma più costoso, ma consente di ottenere una densità di schieramento molto elevata (fino ad alcune decine di migliaia di elementi per singola matrice); tali matrici ad alta densità presentano notevole versatilità sia nello studio dei profili di espressione genica che nell'analisi genotipica: sono particolarmente utili per vedere differenze qualitative nell'espressione genica e per scoprire quali geni sono coinvolti in un determinato processo. Circa un quarto del totale delle pubblicazioni su matrici riguarda questa tipologia.

Ciascuna sonda di DNA è specifica per una singola sequenza complementare, o sequenza bersaglio, di acido nucleico (DNA o RNA) ed è in grado di legarla anche quando questa sia presente in una miscela complessa, quale può essere un estratto cellulare, generando così (quando la sequenza bersaglio sia stata previamente marcata) un segnale misurabile. Ogni specifica sonda della matrice è perciò in grado di rivelare la presenza e, in definite condizioni, di misurare la quantità della molecola complementare previamente marcata presente in un campione biologico.

Quando il bersaglio sia rappresentato dall'mRNA cellulare, l'intensità luminosa di ciascun elemento della matrice fornirà indicazioni sulla quantità di mRNA sintetizzato da ciascuno specifico gene. In altri termini, si otterrà un quadro dell'attività trascrizionale dei singoli geni, ossia un profilo dell'espressione genica cellulare.

Per questo scopo come sonde geniche vengono utilizzate solo sequenze espresse che presentino caratteri di unicità. Ciò significa che tali sequenze, una o più per ciascun gene, devono essere selezionate, nell'ambito dell'intero genoma, tra quelle che vengono trascritte in RNA messaggero, in modo da risultare uniche e specifiche per il gene rappresentato, cioè in modo da non essere presenti in altri geni. Ovviamente ciò è possibile solo attraverso un'analisi mediante computer di una impressionante mole di dati, corrispondenti alla sequenza dell'intero genoma. Le sequenze così selezionate pren-

dono il nome di "sequenze uniche espresse", o EST (Expressed Sequence Tags)[23]. Per lo studio dell'espressione genica, una volta purificati e marcati con fluorocromo gli RNA messaggeri di una cellula o di un tessuto, questi verranno fatti ibridare con le sonde (elementi) della matrice. Ogni sonda, corrispondente a una sequenza unica espressa per un determinato gene, legherà lo specifico mRNA complementare fluorescente. La fluorescenza rilevabile in quel sito (avente precise coordinate, essendo i siti della matrice disposti in righe e colonne) fornirà, in opportune condizioni, una misura dell'mRNA sintetizzato a partire da quel determinato gene e quindi una misura dell'attività di quel gene. La lettura viene effettuata eccitando il marcatore fluorescente con luce coerente (laser) di una determinata lunghezza d'onda e misurando poi la fluorescenza emessa, anch'essa di una specifica lunghezza d'onda. I due marcatori fluorescenti più utilizzati sono il Cy3, che viene eccitato a 550 nm ed emette a 581 nm (luce verde) e il Cy5, che viene eccitato a 649 nm ed emette a 670 nm (luce rossa). I dati di intensità luminosa vengono memorizzati in forma digitale come immagini in formato TIFF (Tagged Image File Format) e poi analizzate al computer utilizzando programmi dedicati. Una delle più comuni applicazioni delle matrici a DNA è quella di confrontare i profili di espressione genica di due diversi campioni biologici, per esempio di un tessuto sano e del corrispondente tessuto patologico, oppure dello stesso tipo di cellule in due differenti condizioni, per esempio in seguito a differenti trattamenti farmacologici. I trascritti dei due campioni da confrontare possono essere marcati con fluorocromi diversi, per esempio Cy3 (verde) e Cy5 (rosso), e poi ibridati a una stessa matrice. Questa procedura viene indicata anche come "schema a due colori". Laddove vi sia maggiore abbondanza di mRNA marcato con Cy3 si avrà fluorescenza verde in corrispondenza del sito ove è presente la sonda complementare, mentre si avrà fluorescenza rossa nel caso opposto in cui sia più abbondante l'mRNA marcato con Cy5. Quando un mRNA marcato con Cy3 e quello corrispondente marcato con Cy5 siano presenti in quantità pressoché uguali, e quindi si leghino in quantità circa equimolari in corrispondenza della sonda specifica, la fluorescenza risulterà gialla. In assenza di entrambi gli mRNA marcati non si avrà alcuna fluorescenza e il sito della sonda apparirà nero. In tal modo l'analisi digitale dell'immagine fornirà precise informazioni sull'espressione genica dei due campioni biologici a confronto (in realtà l'informazione raccolta per ciascun elemento della matrice, cioè per ciascuna sonda, è rappresentata da due valori numerici, corrispondenti alle intensità luminose rilevate dai canali di lettura per Cy3 e per Cy5, che si possono conglobare in un solo numero, corrispondente al rapporto tra tali valori, oppure al logaritmo di tale rapporto). In via alternativa (nello schema a un solo colore), l'immagine ottenuta dopo ibridazione alla matrice di un campione marcato con una determinata sostanza fluorescente può essere confrontata con l'immagine ottenuta dopo ibridazione a una identica matrice di un secondo campione marcato con la stessa sostanza fluorescente. A seconda della matrice utilizzata e del numero di sonde in essa presenti, il confronto potrà riguardare un numero prestabilito di geni o addirittura tutti i geni di una cellula. Gli studi di espressione genica consentono di svelare la complessa rete di regolazione di gruppi di geni in relazione alla loro specifica funzione. Vi sono però casi in cui l'espressione genica non viene studiata per mettere in evidenza il complesso di segnali positivi e negativi che regola l'attività cellulare, ma per riconoscere le caratteristiche distintive di un tipo cellulare che ne consentano la differenziazione da altri tipi. In quest'ultimo caso il profilo di espressione genica viene utilizzato a scopo diagnostico e prognostico per distinguere tra tessuti altrimenti identici quale potrà beneficiare di uno specifico trattamento farmacologico e quale no. Ciò ha importanti implicazioni in ambito clinico, in particolare oncologico, dove, attraverso il peculiare profilo di espressione genica, utilizzato come se si trattasse di una impronta digitale, è possibile classificare un tessuto tumorale in modo da poter impiegare su di esso la terapia più efficace.

Dal momento che vi è una stretta correlazione tra il funzionamento complessivo della cellula e il suo profilo di espressione genica, quest'ultimo ha potuto fornire una formidabile massa di informazioni sui più svariati processi cellulari, dagli stati patologici, degenerativi o neoplastici, all'invecchiamento, alla risposta a farmaci e ormoni e via dicendo. Le matrici a DNA trovano anche impiego in studi di genotipizzazione. In tal caso le sonde della matrice saranno costituite da sequenze genotipospecifiche, che potranno trovare oppure no nel campione biologico saggiato la sequenza ad esse complementare; il quadro di ibridazione risultante definirà in modo univoco il genotipo presente nel campione saggiato.

La specificità di interazione tra sonda e molecola bersaglio è un requisito fondamentale per le matrici. Nel caso delle matrici a DNA la lunghezza minima delle sonde in grado di garantire una sufficiente specificità nelle comuni condizioni di ibridazione è pari a 15-25 nucleotidi.

Come va costruita una matrice

La superficie del supporto al quale devono essere fissati gli elementi della matrice deve rispondere a precisi requisiti. Gli elementi della matrice sono rappresentati da aree di identica forma (circolare) e di uguali dimensioni, tra loro equidistanti e - come sappiamo - disposti in righe e colonne, nonché caratterizzate dalla stessa densità di sonda legata. Per poter ottenere questo risultato, la superficie del supporto deve essere perfettamente planare. Qualsiasi scostamento dalla perfetta planarità può per vari motivi compromettere la qualità dei risultati. La deposizione delle sonde può in-

fatti essere effettuata spruzzando le stesse attraverso una batteria di microscopici ugelli sulla superficie del supporto. Il diametro dell'area circolare corrispondente a ogni singola sonda sarà direttamente proporzionale all'altezza del cono di uscita del getto, ossia alla distanza tra foro di uscita e superficie del supporto. Solo un supporto perfettamente piano potrà quindi garantire che le aree circolari di deposizione delle sonde siano tra loro tutte uguali. Lo scostamento dalla perfetta planarità è causa inoltre di effetti distorsivi anche nella fase di lettura del segnale fluorescente dopo ibridazione, dal momento che le testine del lettore hanno un campo di messa a fuoco molto limitato, cosicché la lettura fatta a distanza non costante tra testina e supporto della matrice può risultare imprecisa.

Oltre a quello della planarità, la superficie del supporto deve rispondere al requisito della omogeneità in termini di reattività chimica. Il trattamento chimico del supporto (con reattivi organosilanici, acrilamide, polilisina, nitrocellulosa) è necessario perché ad esso possano essere successivamente legate le sonde. La densità di gruppi reattivi deve essere la stessa in tutti i punti della matrice, o quanto meno variare di poco da punto a punto. La variazione massima tollerabile nella densità di gruppi reattivi nell'ambito dell'intera superficie trattata è pari a +/- 25%. Il trattamento chimico del supporto è condotto in modo da ottenere una precisa densità dei gruppi reattivi che ad esso si legano: la densità è ottimale quando tali gruppi consentono di legare una quantità massima di sonda in assenza di un eccessivo rumore di fondo. Un aumento della loro densità consente di aumentare l'efficienza di legame con la sonda, ossia la frazione di sonda effettivamente immobilizzata sulla matrice, ma fa parimenti aumentare il legame aspecifico e quindi il rumore di fondo. Un compromesso ottimale si ottiene per concentrazioni di gruppi reattivi capaci di legare una frazione della quantità di sonda compresa tra il 10 e il 30% del totale, per concentrazioni di sonda oligonucleotidica intorno a 30 micromoli/litro o di cDNA intorno a 0.3 microgrammi/microlitro.

Oltre che a supporti di vetro trattato con composti organici di tipo amminico o aldeidico, le sonde possono essere legate a una matrice di acrilamide o di nitrocellulosa. In quest'ultimo caso affinché la reazione di ibridazione tra sonde e molecole bersaglio possa avvenire in maniera ottimale, va tenuto conto del rallentamento della reazione per la minore accessibilità della matrice all'interno della quale le molecole bersaglio devono diffondere per raggiungere le sonde.

Le molecole di sonda non devono staccarsi dalla matrice nel corso dell'esperimento di ibridazione. La stabilità di legame viene considerata sufficiente se nel corso del saggio ibridativo la sonda che si stacca dalla matrice è inferiore al 10% del totale della sonda legata. La possibilità che la sonda si stacchi è dovuta al fatto che durante il saggio sono previste temperature elevate per denaturare il DNA a doppia elica e l'utilizzo di reagenti organici aggressivi, come la formammide nel tampone di ibridazione. L'impiego di reagenti organosilanici per legare le sonde alla superficie di vetro della matrice garantisce notevole stabilità e minima perdita di sonda durante il saggio. Le sonde ancorate attraverso legami non covalenti (come è il caso del DNA che come polianione può essere trattenuto mediante interazione elettrostatica da gruppi cationici fissati alla matrice) hanno invece maggiore tendenza a staccarsi dal supporto.

Concentrazione della sonda negli elementi della matrice

E' intuitivo che quanto maggiore è la densità di sonda (intesa come numero di molecole per singolo elemento della matrice, ovvero numero di molecole per unità di area del singolo elemento), tanto maggiore è la capacità di legame di tali elementi con le molecole bersaglio presenti in un campione biologico. E' altrettanto intuitivo che vi deve essere una densità limite oltre la quale le singole molecole di sonda vengono a essere tra di loro così stipate da non lasciare spazio sufficiente per l'ingresso delle molecole complementari, venendosi a determinare una condizione di impedimento sterico. Per densità di sonda molto basse la capacità ibridativa della sonda è prossima al 100%, ossia virtualmente tutte le molecole di sonda possono legare il bersaglio complementare. All'aumentare della densità di sonda aumenta ovviamente la quantità di bersaglio che può essere legato, ma la percentuale di sonda in grado di ibridare con il bersaglio progressivamente diminuisce. Con procedimenti di titolazione[24] è possibile valutare la quantità di bersaglio legato in funzione della densità di sonda e desumere la densità limite oltre la quale si manifesta l'impedimento sterico; la percentuale di saturazione della sonda è poi immediatamente ottenibile come rapporto tra quantità di sonda presente sull'elemento della matrice e quantità di bersaglio specifico legata. A tale scopo si può per esempio approntare una matrice i cui elementi contengano tutti lo stesso tipo di sonda, ma a differenti densità, essendo stati ottenuti per deposizione sul supporto di concentrazioni della sonda variabili secondo un rapporto scalare comprese in un prefissato intervallo. La matrice viene poi fatta ibridare con una miscela contenente la molecola bersaglio marcata con fluorocromo e alla fine si effettua la lettura in fluorescenza. Ciò che si osserva è che all'aumentare della concentrazione di sonda deposta in ogni singolo sito della matrice si ha un corrispondente aumento della fluorescenza fino a un valore massimo, oltre il quale ogni successivo aumento della concentrazione di sonda comporta una progressiva diminuzione della fluorescenza. Il numero di molecole bersaglio fluorescenti legate dalla sonda in assenza di impedimento sterico viene ad essere infatti direttamente proporzionale alla densità di sonda, ovvero al numero di molecole di sonda per unità di area; con il progressivo aumento di tale densità lo spazio medio tra le

molecole di sonda diminuisce progressivamente e, superato un determinato limite, tende a divenire via via sempre meno accessibile fino a risultare alla fine impenetrabile alle molecole bersaglio marcate. La condizione di impenetrabilità riguarderà tutta l'area circolare del supporto corrispondente al singolo elemento della matrice, ma non la circonferenza dell'elemento, risultando le sonde perimetrali sempre e comunque accessibili alle molecole bersaglio. Per tale motivo all'aumentare della densità di sonda la fluorescenza aumenta progressivamente fino a un valore massimo e poi decresce progressivamente senza però mai annullarsi. La densità di sonda corrispondente alla massima capacità di legame con la molecola bersaglio e quindi al valore massimo di fluorescenza rappresenta la densità ottimale. Per corte sonde oligonucleotidiche a singola elica la densità ottimale, determinabile sperimentalmente come prima descritto, corrisponde a una distanza lineare di circa 20 Angstrom tra le sonde affiancate. Nel caso di corte sonde oligonucleotidiche e per una superficie trattata con gruppi reattivi in modo da ottenere un'efficienza di legame di circa il 30% con i dispositivi per la deposizione delle sonde di normale impiego si ottiene una densità ottimale di sonda utilizzando miscele contenenti la sonda a concentrazioni intorno a 20-30 micromoli/litro. Il vetro non è l'unico supporto per matrici. In teoria è possibile utilizzare altri supporti, come il gel di poliacrilamide[25] o la lamina d'oro[26] o il nylon[27], la nitrocellulosa[28], il polistirene[29] o altri ancora[30,31]. Le sonde possono essere immobilizzate sul supporto sia in forma di singola elica che in forma di doppia elica[26]: quest'ultima può essere rappresentata dalla sonda a singola elica (covalentemente legata per un'estremità a un gruppo chimico per l'ancoraggio al supporto) preibridata con il bersaglio complementare, anch'esso a singola elica. Prima dell'esposizione al bersaglio la matrice con sonde a dsDNA va riportata allo stato di singola elica mediante trattamento termico (lavaggi a temperature intorno a 80 °C).

Altri parametri che condizionano la capacità ibridativa

La densità di sonda non è certamente il solo parametro dal quale dipenda la capacità ibridativa della matrice. Un altro parametro importante è rappresentato dalla competizione di legame tra eliche complementari del DNA. Ciò fa sì che l'efficienza di ibridazione tra la sonda, a singola elica, legata al supporto e la sequenza bersaglio complementare, libera, sia maggiore quando quest'ultima si presenti in forma di singola elica, anziché di doppia elica. Se il bersaglio è a singola elica, questa si può ibridare solo con la singola elica complementare della sonda. Se invece il bersaglio è a doppia elica, vi è competizione di legame tra le due eliche del bersaglio e l'elica, complementare a una di esse, della sonda: in tal caso il legame tra bersaglio e sonda sarà di conseguenza meno efficiente.

Per essere accessibile all'interazione ibridativa, la sonda legata covalentemente per una estremità al supporto deve trovarsi ad una certa distanza da quest'ultimo; per distanze inferiori a un valore critico soglia si verifica una interferenza sterica con lo stesso supporto. Se la catena covalente tra la sonda e il supporto è rappresentata da una sequenza oligonucleotidica, per esempio polidT, si osserva che, superata una certa lunghezza di tale catena, corrispondente a pochi nucleotidi, il segnale ibridativo (es. intensità di fluorescenza) sale oltre lo zero per aumentare gradualmente, fino a un valore massimo, al crescere della lunghezza della catena[24]. Si può quindi affermare che un ulteriore parametro che condiziona l'efficienza di ibridazione è rappresentato dalla lunghezza della catena covalente di ancoraggio della sonda al supporto. Ciò è del resto intuitivo, dal momento che la libertà di movimento (espressa dai gradi di libertà) della sonda aumenta all'aumentare della lunghezza della catena di ancoraggio, del tutto come se si trattasse di un cane legato alla catena. Maggiore libertà di movimento significa anche maggiore probabilità di interagire in condizioni stericamente favorevoli con la catena polinucleotidica complementare e di conseguenza maggiore efficienza di ibridazione.

Caratteristiche del supporto di vetro

Il supporto più utilizzato per le matrici è il vetro, composto principalmente da biossido di silicio SiO_2, o silice, e da percentuali variabili di altri componenti minori (vari ossidi, come quelli di alluminio, boro, calcio, sodio o titanio), che ne modificano entro certi limiti le proprietà fisiche. Il vetro rappresenta un materiale ideale come supporto per le matrici per la sua inerzia chimica, la sua rigidità strutturale, il basso coefficiente di espansione termica, la perfetta trasparenza, ovvero capacità di trasmettere la radiazione elettromagnetica dello spettro visibile, e la bassa fluorescenza intrinseca.

Già si è detto dell'importanza che ha la perfetta regolarità della superficie al fine di ottenere corretti risultati con l'utilizzo delle matrici. I limiti di tolleranza per quanto attiene alle irregolarità della superficie piana del supporto di vetro sono di conseguenza alquanto rigidi. Il vetro deve essere perfettamente levigato e rispondere a determinate specifiche. Le irregolarità possono essere classificate come "graffi" (microscanalature) o avvallamenti: quanto maggiori sono l'ampiezza delle microscanalature e il diametro degli avvallamenti e il loro numero per unità di superficie, tanto maggiore è l'irregolarità (ovvero tanto minore è la regolarità) della superficie del supporto di vetro. La irregolarità della superficie può essere espressa in termini quantitativi attraverso valori numerici che corrispondono all'ampiezza in micrometri delle microscanalature o al diametro degli avvallamenti. Quanto minore è il valore numerico, tanto maggiore è la regolarità della superfi-

cie. Mentre per un normale vetrino da microscopio sono sufficienti specifiche di 80/60, per un supporto per matrice sono necessarie specifiche di 10/5.

Analisi di sequenza e analisi mutazionale con matrici a oligonucleotidi

Anche se l'analisi dell'espressione genica, cioè dell'attività trascrizionale della cellula, rappresenta il principale ambito di utilizzo delle matrici, queste ultime in teoria possono essere progettate e impiegate anche per lo studio di sequenza del DNA, ossia per definire la sequenza di un'unica specie di DNA, previamente amplificata attraverso le consuete tecniche di amplificazione genica, oppure per riconoscere la presenza di una o più mutazioni all'interno di una sequenza genica nota utilizzando come campione una miscela complessa di acidi nucleici. Lo studio quantitativo dell'espressione genica con matrici a cDNA si basa sulla semplice analisi comparativa dell'intensità dei segnali luminosi dopo ibridazione tra le sonde legate (ognuna delle quali presente in eccesso rispetto al proprio bersaglio) e i bersagli marcati che ad esse si legano in modo specifico per complementarietà di sequenza. Si tratta quindi di una analisi concettualmente semplice. L'analisi di sequenza e quella mutazionale di acidi nucleici attraverso l'impiego di matrici a oligonucleotidi è invece concettualmente più complessa e l'analisi dei dati è meno immediata.

Verso la fine degli anni Ottanta e l'inizio dei Novanta sono stati pubblicati i risultati di vari studi condotti allo scopo di ottenere la sequenza ignota di un acido nucleico attraverso la sua interazione di legame con una serie di sonde oligonucleotidiche a sequenza nota[32-34]. Nell'ambito del Progetto di livello mondiale "Genoma Umano", che aveva lo scopo di sequenziare l'intero genoma umano, si cercava di perfezionare vie alternative a quelle classiche di Maxam-Gilbert e di Sanger per lo studio della sequenza del DNA. Con lo sviluppo delle matrici a oligonucleotidi[35,36] il sequenziamento mediante ibridazione ("sequencing by hybridization", o SBH; "sequencing by hybridization with oligonucleotide matrix", o SHOM) è stato reso possibile anche grazie al progressivo sviluppo nel corso degli ultimi quindici anni di algoritmi operativi sempre più validi per l'elaborazione dei dati. Il sequenziamento ibridativo (SBH) si basa sulla possibilità di ottenere una sequenza ignota mediante ibridazione con oligomeri a sequenza nota; le sequenze degli oligomeri che si legano possono essere tra loro confrontate per accertarne il grado di sovrapponibilità. Dalla sovrapposizione di un opportuno numero di sequenze note corrispondenti agli oligomeri che si legano alla sequenza ignota del DNA bersaglio si può in casi ideali ricostruire quest'ultima in modo univoco. Quando un DNA bersaglio a singola elica a sequenza ignota viene analizzato per ibridazione con una matrice a oligonucleotidi si ottiene una informazione su quali sonde (a sequenza nota) della matrice si legano ad esso e sono quindi ad esso perfettamente complementari. In tal modo si viene a individuare uno spettro di sonde complementari (ognuna delle quali ha una sua specifica posizione nella matrice e una sua propria specifica sequenza), senza però ottenere alcuna informazione sul sito di legame di ciascuna sonda lungo la sequenza del DNA bersaglio. La ricostruzione della sequenza bersaglio è tuttavia possibile anche in assenza di questo tipo di informazione, grazie all'utilizzo di particolari algoritmi operativi la cui elaborazione rientra nell'ambito di una nuova branca della biologia molecolare che prende il nome di biologia molecolare computazionale, ma che più propriamente rappresenta sotto tutti gli aspetti una branca della matematica applicata alla biologia[37,38]. L'elaborazione matematica dei dati è infatti assai complessa e può essere fatta utilizzando vari modelli, dai più semplici che ipotizzano l'assenza di errori di appaiamento delle sequenze complementari, ai più complessi che tengono conto della possibilità di appaiamenti errati con un livello di errore che può essere prefissato all'interno di un determinato intervallo di valori. Quanto più basso è il livello di errore ipotizzato, tanto più semplice risulta l'elaborazione matematica e minore la possibilità statistica di errore sul risultato finale.

Analisi quantitativa dell'espressione genica

Una corretta quantificazione di una specie molecolare bersaglio presente nel campione analizzato su matrice richiede che vi sia una relazione di perfetta proporzionalità tra la quantità di molecola bersaglio presente nel campione e la quantità di molecola bersaglio legata dallo specifico elemento della matrice: ciò si traduce in un segnale luminoso di intensità direttamente proporzionale alla quantità di bersaglio. Questa condizione si verifica solo quando la quantità di sonda in ogni elemento della matrice è in eccesso rispetto alla quantità del rispettivo bersaglio presente nel campione biologico da analizzare. Allorché invece la quantità di bersaglio supera la capacità legante della sonda, cosicché quest'ultima viene saturata, vale a dire allorché il bersaglio è in eccesso rispetto alla sonda, si ottiene un segnale luminoso corrispondente a una risposta massimale che non aumenta più di intensità all'aumentare del numero di molecole bersaglio nel campione: in questo caso l'analisi quantitativa non è possibile, poiché la risposta non è più proporzionale, ovvero, come si suol dire, non è più di tipo lineare. La condizione di eccesso di molecola bersaglio va sempre tenuta in considerazione, in particolare quando si analizzino miscele complesse, come quelle di mRNA totale: in queste ultime le differenze quantitative tra le singole specie di mRNA, ossia di mRNA trascritti a partire da geni diversi, possono essere enormi. Gli mRNA più abbondanti possono rappresentare anche l'1% del totale, mentre trascritti più rari possono non raggiungere neppure lo 0,001% del totale. Ciò porta a una saturazione selettiva sulla matrice delle sonde specifiche per le spe-

cie abbondanti, che così non possono essere correttamente quantificate, mentre le sonde specifiche per i trascritti meno abbondanti si possono mantenere in eccesso, il che permette di conservare le condizioni di linearità nella misura. In condizioni di saturazione selettiva delle sonde i dati ottenibili risultano essere quantitativi per le specie poco rappresentate e selettivamente non quantitativi per le specie più abbondanti. Questo effetto, noto come "compressione del segnale", può portare a una lettura erronea sulle fasce alte di concentrazione, ciò che si cerca di evitare in tutte escluse poche situazioni sperimentali. In alcune situazioni l'eccesso di bersaglio viene esplicitamente ricercato, per esempio per misurare la densità assoluta di sonda sulla matrice. In questo tipo di analisi si cerca di ottenere la saturazione completa della sonda, corrispondente alla concentrazione di bersaglio specifico oltre la quale non si verifica alcun aumento nell'intensità del segnale luminoso; confrontando quest'ultimo con quello di una quantità nota di marcatore fluorescente è possibile calcolare il numero assoluto di molecole sonda per unità di area nell'elemento di matrice, ossia la densità di sonda. Una condizione di saturazione selettiva delle sonde può rendersi necessaria per misurare la concentrazione di specie rare in miscele complesse, come i trascritti di geni poco espressi. In questi casi la matrice verrà cimentata con forti quantità di mRNA totale, tali da saturare le sonde per i trascritti più abbondanti, fornendo al contempo quantità rivelabili di trascritti rari, che, non saturando le corrispondenti sonde sulla matrice, possono così venire correttamente quantificati. Nelle condizioni in cui si ha una lettura lineare per i trascritti più rappresentati i trascritti più rari potrebbero infatti generare un segnale eccessivamente debole o non rilevabile.

Per poter confrontare, dopo averli memorizzati in forma digitale, i dati ottenuti da campioni biologici diversi e da esperimenti differenti, occorre disporre di segnali la cui intensità possa essere presa come termine di riferimento alla quale rapportare tutte le altre. Come riferimento viene spesso utilizzato il segnale corrispondente ai cosiddetti "geni costitutivi" (housekeeping genes)[39], ossia a quei geni che, esplicando un ruolo centrale nel metabolismo cellulare, sono espressi all'incirca in egual modo in tutte le cellule e tessuti. Va tuttavia tenuto in considerazione il fatto che l'invarianza nell'espressione dei geni costitutivi vale solo entro certi limiti, poiché in alcune condizioni il loro livello di espressione si può modificare[39,40]. Una alternativa più rigorosa è allora quella di costruire delle curve di calibrazione aggiungendo al materiale oggetto di analisi delle quantità note di specifici mRNA non presenti nel campione da analizzare utilizzati come controllo. Le sonde specifiche sulla matrice, che dovranno ovviamente essere presenti in eccesso rispetto al bersaglio, legheranno quest'ultimo, in quantità nota, e le corrispondenti fluorescenze potranno essere utilizzate come termini di paragone: si otterrà una serie di segnali corrispondenti a quantità note di mRNA. Ciò permette una quantificazione sia assoluta che relativa. La curva standard, o di calibrazione, può servire anche a correggere l'effetto di compressione del segnale che si verifica in caso di eccesso di molecola bersaglio rispetto alla sonda.

Normalizzazione e analisi dei dati

Il confronto tra dati ottenuti da una serie di matrici è possibile, come si è testé riferito, solo quando in queste siano presenti identici elementi di riferimento. In altri termini, è necessario avere, almeno in linea teorica e idealmente, misure ottenute tutte con lo stesso metro. Per poter procedere alla loro elaborazione, è necessario che i dati vengano corretti per le distorsioni che essi possono subire a causa di condizioni sperimentali non perfettamente omogenee[41-43]. Le operazioni che hanno lo scopo di rimuovere tali distorsioni vengono definite di normalizzazione[44,45]. La normalizzazione dei dati ottenuti da una serie di matrici è assai più semplice e rigorosa quando vi siano punti di riferimento identici a cui rapportare i valori di lettura. Come sopra riportato, ciò è ottenibile utilizzando per tutte le matrici un comune riferimento quantitativo, rappresentato dai geni costitutivi, o meglio ancora da quantità note di acido nucleico estraneo al campione in esame in modo da ottenere una curva di riferimento per la normalizzazione. Permane comunque la possibilità di normalizzare, ovvero rendere tra loro compatibili e confrontabili, risultati ottenuti in esperimenti differenti in cui i riferimenti utilizzati non siano gli stessi. La normalizzazione così ottenibile offre necessariamente minore precisione di quella basata su controlli identici. Esistono numerosi metodi finalizzati a eliminare o ridurre anche altri errori sistematici che si possono verificare nell'utilizzo delle matrici a sonde. I sistemi di lettura sono controllati da programmi in grado di valutare la qualità delle immagini ottenute. Ciò consente di escludere immagini anomale, che non rispettino i parametri previsti. E' inoltre possibile sottrarre il rumore di fondo, il che garantisce una maggiore purezza dei dati. Per le matrici a due colori i dati numerici relativi alle intensità luminose misurate per Cy3 e per Cy5 vengono espressi come logaritmo in base 2 del rapporto tra le intensità. Questa conversione rende più facile le successive analisi, perché valori del logaritmo pari a zero (corrispondenti a valori del rapporto tra le due intensità per Cy3 e per Cy5 pari a 1) indicano che non vi è alcuna variazione nell'espressione genica, valori positivi del logaritmo corrispondono invece ad un aumento dell'espressione genica e valori negativi a una sua diminuzione. Sempre per le matrici a due colori è prevista una normalizzazione intra-matrice in grado di correggere errori dovuti a disomogeneità nella marcatura del bersaglio con Cy3 e Cy5 e a vizi di lettura del fotomoltiplicatore, che potrebbe mostrare efficienze diverse sui due canali. I metodi correttivi contemplano la regres-

sione lineare di Cy5 contro Cy3, la regressione lineare di $\log_2$(Cy5/Cy3) contro intensità media o in altri casi la regressione non lineare (di Loess) dello stesso logaritmo contro intensità media[46-48]. Per rendere confrontabili i dati ottenuti con matrici diverse, a uno o due colori che siano, si possono poi utilizzare metodi che contemplano l'allineamento delle medie (o in alcuni casi delle mediane) dei valori ottenuti per ogni singola matrice, oppure il più usato metodo della "centratura" dei dati, che si basa sul presupposto che media e deviazione standard dei valori ottenuti per le singole matrici debbano essere sempre le stesse, oppure ancora la normalizzazione della distribuzione, metodo più complesso e meno usato, utile quando i gruppi di dati da analizzare presentino distribuzioni che si discostino dalla normale gaussiana.

Una volta utilizzate tutte le più opportune procedure atte a rendere tra loro confrontabili i dati sperimentalmente ottenuti, questi possono essere successivamente analizzati per estrapolare l'informazione biologica in essi racchiusa. La loro successiva elaborazione rappresenta l'oggetto di studio di una nuova branca della matematica applicata, la bioinformatica. Si possono a tale scopo utilizzare svariati programmi, peraltro in continua evoluzione, che permettono di tradurre l'informazione numerica nella forma più utile per fornire una risposta agli specifici quesiti biologici di volta in volta posti. Un primo e più semplice livello di analisi riguarda la significatività statistica del risultato[49,50]. Si supponga di voler identificare quali geni sono stimolati o inibiti in seguito a un determinato trattamento farmacologico. In tal caso si andranno, per esempio, a confrontare i quadri di espressione genica di un certo numero di soggetti prima e dopo il trattamento. L'analisi statistica andrà condotta su ogni singolo gene, per stabilire se la sua espressione nel gruppo di soggetti trattati sia significativamente diversa rispetto a quella osservata nel gruppo degli stessi soggetti prima del trattamento. Questi tipo di analisi dovrà essere condotta in parallelo su tutti i geni rappresentati sulla matrice. Si partirà dall'ipotesi nulla, e cioè dall'ipotesi che la variabilità dei risultati non sia riconducibile a una reale differenza nell'espressione genica. Per ogni singolo gene si confronteranno i valori relativi a due gruppi, quello prima e quello dopo il trattamento. Si potrà calcolare il valore del rapporto critico di *t*, che dipende da media, deviazione standard e numero di misure ottenute per ciascun gruppo. Confrontando il valore di *t* con la curva di distribuzione del *t* di Student relativa ai gradi di libertà in gioco, si potrà ottenere il valore di *p*, ossia della probabilità statistica che la differenza misurata sia compatibile con l'ipotesi nulla. Ovviamente, quanto più piccolo risulterà il valore di *p*, tanto minore sarà la probabilità che la differenza osservata sia imputabile al caso. In realtà questa classica procedura vale solo in prima approssimazione, in quanto l'analisi, condotta per fortuna attraverso software dedicati in grado di reiterarla per ciascun singolo gene rappresentato sulla matrice, è in generale più complessa e si avvale preferibilmente di metodi statistici non parametrici, che garantiscono una maggiore precisione quando, come in questi casi, la distribuzione dei dati (che, riguardando un singolo gene per volta, sono in numero ridotto) si discosta assai spesso da quella normale gaussiana.

Un successivo livello di analisi, sempre reso possibile dall'elaborazione elettronica mediante programmi dedicati[51-54], è quello inteso a evidenziare relazioni tra dati, che potrebbero sottendere effettive relazioni di carattere biologico. Geni la cui espressione risulta simultaneamente aumentata o diminuita potrebbero avere ruoli funzionali simili all'interno della cellula o appartenere a vie metaboliche correlate. Mentre l'analisi di significatività statistica dei risultati riguarda i singoli geni e viene condotta in parallelo su ciascuno di essi, l'analisi finalizzata a evidenziare relazioni tra geni e a identificare geni che si comportano in maniera coordinata deve necessariamente essere condotta confrontando tra loro, in linea teorica simultaneamente, tutti i geni rappresentati sulla matrice. I dati numerici raccolti per un numero *n* di geni in una serie di *m* condizioni sperimentali diverse (o campioni diversi) possono essere disposti all'interno di una matrice numerica *n* x *m*, ossia di *m* colonne e *n* righe, che si potrebbe definire una matrice dei dati complessivi di espressione genica, dove nelle *n* righe (orizzontali) sono riportati i valori relativi a ogni singolo gene nelle *m* diverse condizioni sperimentali (ovvero negli *m* campioni) e nelle *m* colonne (verticali) i valori ottenuti in ogni singola condizione sperimentale (o singolo campione) su tutti gli *n* geni. Per ciascuno degli *n* geni l'informazione raccolta, corrispondente ai dati di ciascuna riga della matrice di espressione *n* x *m*, si può esprimere in forma vettoriale, ossia sotto forma di un "vettore di espressione". Tale vettore rappresenta l'espressione di un gene attraverso una successione di esperimenti. L'analisi di correlazione tra geni diversi è allora matematicamente traducibile nell'analisi di distanza tra i corrispondenti vettori di espressione. Ognuno di questi possiede *m* coordinate cartesiane ed è rappresentabile in uno spazio a *m* dimensioni. Ciascun gene, corrispondente a ciascuna singola riga sulla matrice dei dati, può essere considerato come un punto in uno spazio *m*-dimensionale, dove *m* è il numero delle colonne, ognuna delle quali corrispondente a una singola condizione sperimentale. Analogamente, i dati corrispondenti a ciascuna condizione sperimentale (ovvero ciascun campione analizzato) si possono riassumere in un vettore a *n* coordinate, ovvero rappresentabile in uno spazio a *n* dimensioni, dove *n* è il numero dei geni analizzati, ognuno dei quali corrispondente a una singola riga della matrice dei dati. Se il confronto tra geni si traduce nell'analisi della distanza tra i rispettivi vettori di espressione (in uno spazio a *m* dimensioni), lo stesso vale per il confronto tra campioni (o condizioni sperimentali), dove la distanza tra vettori va misurata in uno spazio a *n* dimensioni. In Tabella I è rappresentata la matrice dei dati di espressione genica per i tre geni A,

Tabella I. Matrice di espressione di tre geni A, B, C in due diverse condizioni sperimentali X e Y. Con i simboli x1, x2, x3, y1, y2, y3 sono indicati i valori di espressione genica ottenuti dopo lettura in fluorescenza.

	X	Y
A	x1	y1
B	x2	y2
C	x3	y3

B, C in due diverse condizioni (o due diversi campioni) X e Y. Si tratta di una matrice *n* x *m*, ossia a *m* colonne e *n* righe, dove $m = 2$ e $n = 3$. Il confronto tra A, B e C (Fig. 2a) implica la valutazione della distanza tra i rispettivi vettori (definiti in uno spazio a $m = 2$ dimensioni, essendo 2 i valori numerici per ciascuna riga della matrice), la cui origine coincide con quella degli assi cartesiani ortogonali e la cui estremità è data rispettivamente dai punti A, B, C aventi coordinate: $A(x_1, y_1)$; $B(x_2, y_2)$; $C(x_3, y_3)$. Il confronto tra X e Y (Fig. 2b) richiede la valutazione della distanza tra i corrispondenti vettori (definiti in uno spazio a $n = 3$ dimensioni, essendo 3 i valori numerici per ciascuna colonna della matrice), la cui origine coincide con quella degli assi cartesiani ortogonali e la cui estremità è data rispettivamente dai punti X e Y di coordinate: $X(x_1, x_2, x_3)$; $Y(y_1, y_2, y_3)$. Applicando semplicemente il teorema di Pitagora, si può dimostrare che la distanza tra A e B, aventi rispettivamente coordinate cartesiane (x_1, y_1) e (x_2, y_2), è pari alla radice quadrata della somma $(x_1 - x_2)^2 + (y_1 - y_2)^2$. Con considerazioni analoghe è possibile calcolare la distanza, che prende il nome di euclidea, tra profili di espressione contenenti un numero *k* di valori, ovvero tra vettori di espressione in uno spazio a *k* dimensioni. La somiglianza tra i profili di espressione di due geni A e B, oppure di due campioni, oltre che come distanza euclidea, può essere espressa come coefficiente di correlazione *r*, cui corrisponde una formula complessa che non serve qui ricordare, e il cui valore varia nell'intervallo compreso tra -1 e +1. Un valore di $r = -1$ indica una forte correlazione negativa, il che significa che quando il gene A è fortemente espresso, l'espressione di B è estremamente bassa. Un valore $r = +1$ indica invece una forte correlazione positiva: l'attività di entrambi i geni varia in maniera del tutto simile. Un valore di *r* pari a zero indica che tra A e B non esiste alcuna correlazione. La scelta di usare la distanza euclidea piuttosto che il coefficiente di correlazione *r* (quest'ultimo utilizzato nei metodi di Spearman o di Pearson)[55] come parametri per valutare il grado di somiglianza tra profili di espressione genica non è indifferente ai fini del risultato e va ponderata in relazione al tipo di protocollo operativo impiegato. Una volta stabilite le regole matematiche per poter confrontare tra loro i vari profili di espressione genica, è possibile applicarle per raggruppare i profili in base al grado di

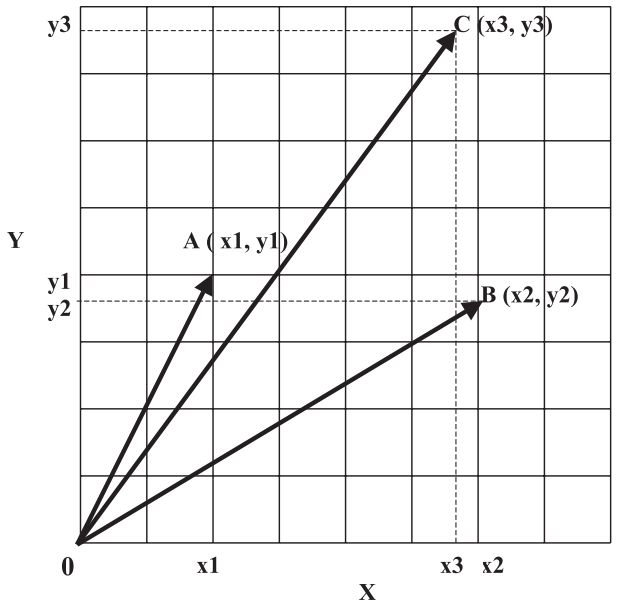

Figura 2a. Rappresentazione vettoriale dei tre geni A, B, C nelle due condizioni sperimentali (o nei due campioni) X e Y.

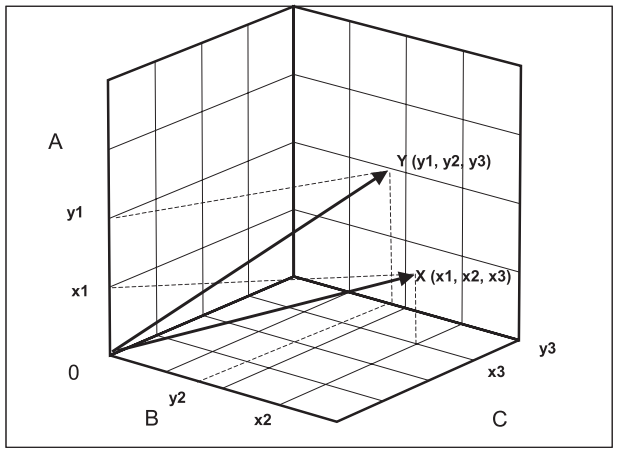

Figura 2b. Rappresentazione vettoriale delle due condizioni sperimentali (o dei due campioni) X e Y attraverso i rispettivi valori di espressione genica (per i geni A, B, C).

somiglianza. Esistono vari metodi di raggruppamento dei dati, ideati per estrarre da questi ultimi l'informazione biologica essenziale[56-64]. I metodi "agglomerativi" iniziano l'analisi dai singoli elementi e procedono associando tra loro uno dopo l'altro quelli che presentano il maggior grado di correlazione, fino a esaurire tutti i profili della matrice dei dati. I metodi "divisivi" o "partitivi" seguono un iter opposto, partendo dal complesso dei profili, che vengono dapprima suddivisi in due o più sottogruppi comprendenti ognuno profili tra loro simili; l'operazione viene ripetuta su ciascun sottogruppo, che viene ulteriormente suddiviso, e così di seguito, fino ad arrivare ai singoli profili. La tecnica di analisi dei profili più largamente usata è quella del "raggruppamento (clustering) gerarchico", che

consente di costruire un dendrogramma (una sorta di albero genealogico o filogenetico) dei geni o dei campioni, nel quale gli elementi più simili vengono a trovarsi in raggruppamenti tra loro vicini e quelli meno simili in raggruppamenti tra loro lontani. Questo tipo di analisi, la cui teoria matematica risale in buona parte agli anni Trenta, cominciò a essere applicato alle matrici per espressione genica verso la fine degli anni Novanta[65]. La correlazione tra i vari geni espressa in forma di dendrogramma risulta assai familiare, comprensibile e gradita ai genetisti. Il procedimento generalmente usato per il raggruppamento gerarchico è quello di tipo agglomerativo. Ciascuno degli *n* profili genici può essere confrontato con se stesso e con gli altri *n*-1 profili. Le distanze relative tra un profilo e l'altro (ovviamente ottenute tutte con uno stesso metodo, ossia tutte come distanze euclidee tra vettori, oppure, per esempio, tutte con il metodo di correlazione di Spearman) possono essere riportate in una matrice numerica *n* x *n* delle distanze. Ovviamente la distanza tra ciascun profilo e se stesso sarà sempre pari a zero. Sulla matrice delle distanze si individuano i due profili che presentano la distanza più piccola tra loro. Tali profili vengono raggruppati insieme. Si ricalcola la distanza tra questo gruppo e tutti gli altri gruppi o singoli profili. Esistono modi alternativi per calcolare tale distanza: in base al metodo "single linkage" la distanza tra due gruppi A e B è uguale alla più piccola distanza tra qualsiasi elemento di A e qualsiasi elemento di B; in base al metodo "complete linkage" la distanza tra i due gruppi è all'opposto definita come la massima distanza tra qualsiasi elemento di A e qualsiasi elemento di B; con il metodo "average linkage" la distanza tra A e B è invece data dalla distanza media tra tutti gli elementi presenti in ciascun gruppo. Una volta ricalcolata con uno di questi metodi la distanza, si individuano i due gruppi o profili che presentano la distanza più piccola tra loro: questi vengono raggruppati insieme. L'operazione procede reiterativamente, producendo una serie di gruppi sempre più grandi tra loro interconnessi, nonché una tabella di distanze relative: viene così generato un dendrogramma Gli elementi con i quadri di espressione più simili saranno tra loro collegati attraverso tratti brevi, mentre i gruppi associati nei passaggi via via successivi saranno tra loro connessi da tratti sempre più lunghi. L'altezza dei tratti di interconnessione è direttamente proporzionale ai valori di distanza relativa, che possono anche essere riportati a fianco di ciascun tratto. Nello schema a dendrogramma di solito i singoli valori di espressione genica vengono rappresentati all'interno della matrice dei dati attraverso una scala di colori convenzionali, che forniscono un'informazione visiva immediata dell'attività trascrizionale dei geni. Un gradiente di colore che va, per esempio, dal nero al rosso rappresenta valori del logaritmo del rapporto di espressione che a partire dallo zero (corrispondente a un valore pari a 1 per il rapporto, ossia a nessuna variazione nell'intensità di espressione genica) salgono verso valori positivi crescenti. Il colore rosso denota in tal caso una iperespressione del gene. Un gradiente di colore che va, per esempio, dal nero verso il verde rappresenta invece logaritmi del rapporto di espressione che a partire dallo zero (espressione invariata) scendono verso valori negativi. Il colore verde denota in tal caso una ipoespressione, ovvero un'inibizione dell'espressione genica (Fig. 3).

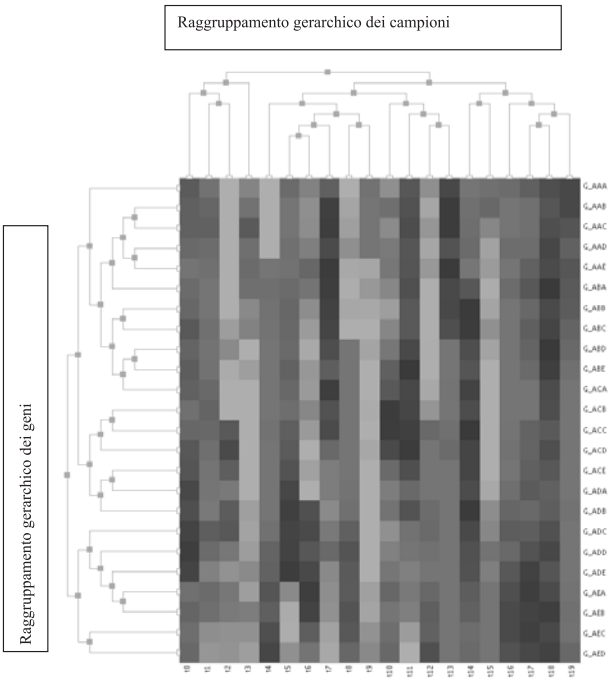

Figura 3. Analisi per raggruppamento gerarchico dei geni e dei campioni. Il dendrogramma di sinistra si riferisce ai geni, quello in alto ai campioni. All'interno della matrice dei dati i valori di espressione genica sono rappresentati attraverso una scala di colori convenzionali che forniscono un'informazione visiva immediata dell'attività trascrizionale dei geni.

Applicazioni cliniche delle matrici

Come si è finora ripetuto, l'applicazione principale delle matrici a sonde nell'ambito della ricerca è rappresentato dall'analisi dell'espressione genica, che ha forti ricadute anche in ambito clinicodiagnostico. Per fare un esempio, l'alterazione della regolazione trascrizionale nella cellula cancerosa dà luogo a profili trascrizionali che, in opportune condizioni operative, possono risultare patognomonici e fornire indicazioni di estrema utilità per il clinico. E' possibile identificare e caratterizzare i geni la cui espressione è più fortemente alterata in un determinato processo patologico, di tipo neoplastico, infiammatorio o degenerativo, per esempio, ottenendo così di individuare marcatori biologici di utilità diagnostica in casi specifici[66-71]. L'analisi molecolare mediante matrici a sonde consente in ambito oncologico una fine classificazione della neoplasia, con importanti implicazioni di carattere prognostico e te-

rapeutico[72,73], come l'identificazione di bersagli molecolari per la chemioterapia. Si va sempre più estendendo l'uso di farmaci diretti contro specifici bersagli molecolari, come è il caso degli oligonucleotidi antisenso, degli RNA interferenti, o "silenziatori" (siRNAs), o di altri inibitori molecolari, come gli "small molecule inhibitors" (SMIs): l'efficacia di tali farmaci nel singolo paziente neoplastico dipende dal livello di espressione e dall'importanza funzionale dello specifico bersaglio molecolare nel tessuto neoplastico di quel paziente. Il profilo di espressione genica può essere utilizzato come una sorta di impronta digitale per caratterizzare la neoplasia e indirizzare verso il trattamento chemioterapico più efficace. L'analisi del profilo di espressione genica consente di poter prevedere la risposta del singolo paziente agli agenti citotossici utilizzati nella chemioterapia dei tumori, fornendo informazioni utili per un loro impiego ottimale e personalizzato, cioè tagliato su misura del singolo e inteso a conseguire la massima efficacia con il minimo di effetti tossici collaterali[72]. Lo studio degli effetti dei farmaci sulla cellula condotto mediante analisi su matrici ha stimolato lo sviluppo di una nuova branca della farmacologia, che prende il nome di "farmacogenomica" (che diventa "tossicogenomica" quando lo studio verta in particolare sugli effetti tossici)[74-79]. Le matrici rappresentano uno strumento unico e insostituibile per lo studio dell'impatto dei farmaci sul metabolismo cellulare. L'omeostasi cellulare è infatti controllata da una complessa rete genica: l'attività di centinaia di geni può essere modulata, cioè stimolata o inibita, in risposta a un singolo effettore. Lo studio dell'effetto dei farmaci sull'espressione genica consente di ottenere informazioni di estrema utilità per individuarne il meccanismo d'azione e valutarne l'utilità clinica in relazione ai polimorfismi genetici che condizionano la risposta (in termini di efficacia e di tossicità) nel singolo paziente[80]. Studiando le risposte nei soggetti trattati è possibile individuare dei profili di espressione genica correlati con una maggiore o minore efficacia o con una maggiore o minore tossicità di un farmaco, che permettono, previa analisi su matrice, di valutare per un singolo paziente quale tra le terapie proponibili sia la migliore (la più efficace e meno tossica). Una volta studiati con matrici ad ampio spettro di sonde i profili generali di espressione genica di specifici tipi di tumore e una volta analizzata la loro associazione con una maggiore o minore sensibilità verso una determinata terapia o con una migliore o peggiore prognosi, è possibile estrapolare tra tutte le sonde utilizzate quelle di maggiore interesse ai fini del profilo diagnostico. Queste potranno essere impiegate per costruire matrici dedicate (a basso numero di elementi), utilizzabili solo per scopi specifici e ben delimitati, particolarmente utili nella diagnostica clinica. La riproducibilità dei risultati è condizione necessaria per un corretto impiego delle matrici a sonde e ciò vale anche e soprattutto per il loro utilizzo in ambito clinicodiagnostico. Gli studi che vertono sul controllo di qualità sono appunto intesi a verificare la riproducibilità dei risultati, sia quando uno specifico campione sia analizzato più volte con uno stesso tipo di matrice, sia quando esso venga analizzato con matrici differenti[81-86]. Matrici a basso numero di elementi e perciò di costo limitato vengono sempre più utilizzate nella diagnostica clinica per l'identificazione e la genotipizzazione di specifici tratti genomici; esse consentono di evidenziare la presenza di specifici agenti infettivi, evidenziandone caratteristiche peculiari importanti a fini diagnostici e terapeutici (per esempio con una sola analisi è possibile la genotipizzazione dei papillomavirus umani presenti in un campione biologico e la valutazione della loro attività trascrizionale a effetto oncogeno); allo stesso modo è possibile l'analisi fine dei polimorfismi genetici associati a specifiche patologie e più in generale la ricerca di malattie genetiche. E' evidente che le matrici ad elevato numero di sonde continueranno a trovare impiego nella ricerca biologica di base e che quelle a basso numero di elementi potranno essere progettate per analisi in un ambito più circoscritto, avendo bersagli di volta in volta diversi, che potranno essere acidi nucleici da cellule umane, piuttosto che di provenienza batterica o virale. Quanto detto finora riguarda le matrici a DNA; anche le matrici a proteine trovano però importanti applicazioni cliniche, in particolare nell'ambito oncologico, allergologico e delle patologie autoimmuni[87-90]. Al supporto solido della matrice possono venire legati, tra l'altro, antigeni o anticorpi. Le matrici ad antigeni consentono di legare le molecole anticorpali presenti nel campione biologico da analizzare e di ottenere profili di risposta anticorpale utili, in ben definiti contesti, per stabilire la diagnosi e indirizzare la terapia. Utilizzando matrici ad allergeni è possibile determinare il preciso profilo di risposta di tipo IgE e delineare il quadro completo di allergia del paziente. Con un singolo passaggio analitico è possibile studiare la risposta di un soggetto verso migliaia di differenti allergeni ed epitopi[91,92]. Non molto diverso concettualmente risulta l'utilizzo di matrici ad antigeni "tumore-associati" per la diagnosi sierologica di specifiche neoplasie. Le cellule neoplastiche vengono riconosciute dal sistema immunitario come estranee ("not self") e inducono la produzione di autoanticorpi contro proteine cellulari autologhe, indicate come "antigeni associati ai tumori". Questi autoanticorpi possono rappresentare un utile strumento sierologico, dal momento che ciascun tipo di tumore appare contrassegnato da una espressione di tipo autoanticorpale caratteristica, che può essere sfruttata a fini diagnostici. Per questo motivo le matrici ad antigeni tumore-associati promettono di diventare un valido strumento per la diagnosi precoce, il monitoraggio della progressione del tumore, la valutazione prognostica, la definizione di una terapia ottimale, il monitoraggio della terapia stessa, nonché l'identificazione di nuovi bersagli della terapia[93-95]. Più in generale, l'impiego di matrici ad antigeni consente di delineare il profilo di risposta anticorpale (che può es-

sere altamente specifico) verso autoantigeni, nelle malattie autoimmuni, verso antigeni tumorali, come pure verso antigeni microbici[96-100]. L'individuazione di specifici quadri anticorpali può essere di grande aiuto per una diagnosi precoce delle malattie autoimmuni, quando ancora i segni clinici non si sono ancora manifestati e nella classificazione prognostica in base alla velocità di progressione della malattia. Anche matrici ad anticorpi possono essere utilizzate nello studio dei tumori: esse consentono di evidenziare la presenza dei rispettivi antigeni sia liberi nel siero, nel plasma e in altri fluidi corporei, sia legati a membrane cellulari. Il quadro antigenico può fornire indicazioni di carattere diagnostico o prognostico[101,102]. E' possibile, per esempio, una caratterizzazione immunofenotipica delle cellule leucemiche di indiscutibile utilità clinica[103]. Com'è ovvio, si possono qui definire solo nelle loro linee essenziali le applicazioni cliniche delle matrici, invitando necessariamente coloro che fossero motivati ad approfondire l'argomento alla consultazione della letteratura specifica, prendendo spunto dai lavori citati.

Conclusione

L'introduzione delle matrici nella diagnostica clinica è assai recente ed è lecito pensare che questa tecnologia stia movendo solo ora i primi passi. Il costo delle matrici è elevato, tanto più elevato quanto maggiore è il numero di sonde richieste. Va tuttavia considerato che per utilizzi specifici, come si è ribadito, il numero di sonde può essere ridotto, contenendo la spesa. La costruzione di una matrice somiglia sotto molti aspetti a quella dei "microchip" elettronici e si avvale di tecnologie non molto differenti. Se si considerano l'evoluzione tecnologica e l'andamento dei prezzi degli elaboratori elettronici negli ultimi anni, ragionando per analogia si è tentati di pronosticare un futuro non troppo dissimile per le matrici a sonde. Come si è verificato nel recente passato per altre tecnologie diagnostiche innovative, dall'ambito della ricerca si è giunti in tempi brevi all'utilizzo nel laboratorio clinico. E' plausibile che le matrici non debbano fare eccezione.

Bibliografia

1. Grunstein M, Hogness DS. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc Natl Acad Sci USA 1975; 72:3961-5.
2. Benton WD, Davis RW. Screening lambdagt recombinant clones by hybridisation to single plaques in situ. Science 1977; 196:180-2.
3. St. John TP, Davis RW. Isolation of galactose-inducible DNA sequences from Saccharomyces cerevisiae by differential plaque filter hybridisation. Cell 1979; 16:443-52.
4. Khrapko KR, Lysov YP, Khorlyn AA, Shick VV, Florentiev VL, Mirzabekov AD. An oligonucleotide hybridisation approach to DNA sequencing. FEBS Lett 1989; 256:118-22.
5. Krapko KR, Lysov YP, Khorlin AA, Ivanov IB, Yershov GM, Vasilenko SK, et al. A method for DNA sequencing by hybridisation with oligonucleotide matrix. DNA Seq 1991; 1:375-88.
6. Pevzner PA, Lysov YP, Khrapko KR, Beyavsky AV, Florentiev VL, Mirzabekov AD. Improved chips for sequencing by hybridisation. J Biomol Struct Dyn 1991; 9: 399-410.
7. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spacially addressable parallel chemical synthesis. Science 1991; 251:767-73.
8. Maskos U, Southern EM. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthetised in situ. Nucleic Acid Res 1992; 20:1679-84.
9. Maskos U, Southern EM. A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesised on a glass support. Nucleic Acids Res 1993; 21: 4663-9.
10. Nizetic D, Zehetner G, Monaco AP, Gellen L, Young BD, Lehrach H. Construction, arraying, and high-density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries. Proc Natl Acad Sci USA 1991; 88:3233-7.
11. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270:467-70.
12. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10614-9.
13. Cekaite L, Hovig E, Sioud M. Protein arrays: a versatile toolbox for target identification and monitoring of patient immune responses. Methods Mol Biol 2007; 360: 335-48.
14. Kricka LJ, Master SR, Joos TO, Fortina P. Current perspectives in protein array technology. Ann Clin Biochem 2006; 43:457-67.
15. Yu X, Xu D, Cheng Q. Label-free detection methods for protein microarrays. Proteomics 2006; 6:5493-503.
16. Becker KF, Metzger V, Hipp S, Hofler H. Clinical proteomics: new trends for protein microarrays. Curr Med Chem 2006; 13:1831-7.
17. Hewitt SM. The application of tissue microarrays in the validation of microarray results. Methods Enzymol 2006; 410:400-15.
18. Eguiluz C, Viguera E, Millan L, Perez J. Multitissue array review: a chronological description of tissue array techniques, applications and procedures. Pathol Res Pract 2006; 202:561-8.
19. JubbAM, Pham TQ, Frantz GD, Peale FV Jr, Hillan KJ. Quantitative in situ hybridisation of tissue microarrays. Methods Mol Biol 2006; 326:255-64.
20. Braunschweig T, Chung JY, Hewitt SM. Tissue microarrays: bridging the gap between research and the clinic. Expert Rev Proteomics 2005; 2:325-36.
21. Watanabe A, Cornelison R, Hostetter G. Tissue microarrays: applications in genomic research. Expert Rev Mol Diagn 2005; 5:171-81.
22. Fedor HL, Marzo AM. Practical methods for tissue microarray construction. Methods Mol Med 2005; 103:89-101.
23. Boguski MS, Lowe TM, Tolstoshev CM. dbEST – database for "expressed sequence tags". Nat Genet 1993; 4: 332-3.

24. Guo Z, Guilfoyle RA, Thiel AJ, Wang R, Smith LM. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Res 1994; 22:5456-65.
25. Wang Y, Wang H, Gao L, Liu H, Lu Z, He N. Polyacrylamide gel film immobilized molecular beacon array for single nucleotide mismatch detection. J Nanosci Nanotechnol 2005; 5:653-8.
26. Peterson AW, Heaton RJ, Georgiadis RM. The effect of surface probe density on DNA hybridization. Nucleic Acids Res 2001; 29:5163-8.
27. Van Ness J, Kalbfleisch S, Petrie CR, Reed MW, Tabone JC, Vermeulen NMJ. A versatile solid support system for oligodeoxynucleotide probe-based hybridisation assays. Nucleic Acids Res 1991; 19:3345-50.
28. Saiki RK, Walsh PS, Levenson CH, Erlich HA. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. Proc Natl Acad Sci USA 1989; 86:6230-4.
29. Rasmussen SR, Larsen MR, Rasmussen SE. Covalent immobilization of DNA onto polystyrene microwells: the molecules are only bound at the 5' end. Anal Biochem 1991; 198:138-42.
30. Gingeras TR, Kwoh DY, Davis GR. Hybridization properties of immobilized nucleic acids. Nucl Acids Res 1987; 15:5373-90.
31. Lund V, Schmid R, Rickwood D, Hornes E. Assessment of methods for covalent binding of nucleic acids to magnetic beads, Dynabeads™, and the characteristics of the bound nucleic acids in hybridization reactions. Nucl Acids Res 1988; 16:10861-80.
32. Khrapko KR, Lysov YP, Khorlyn AA, Shick VV, Florentiev VL, Mirzabekov AD. An oligonucleotide hybridization approach to DNA sequencing. FEBS Lett 1989; 256:118-22.
33. Drmanac R, Labat I, Brukner I, Crkvenjakov R. Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics 1989; 4:114-28.
34. Strezoska Z, Paunesku T, Radosavljevic, Labat I, Drmanac R, Crkvenjakov R. DNA sequencing by hybridization: 100 bases read by a non-gel-based method. Proc Natl Acad Sci USA 1991; 88:10089-93.
35. Pevzner PA, Lysov YP, Khrapko KR, Belyavsky AV, Florentiev VL, Mirzabekov AD. Improved chips for sequencing by hybridization. J Biomol Struct Dyn 1991; 9: 399-410.
36. Gunderson KL, Huang XC, Morris MS, Lipshutz RJ, Lockhart DJ, Chee MS. Mutation detection by ligation to complete *n*-mer DNA arrays. Genome Res 1998; 8: 1142-53.
37. Zhang JH, Wu LY, Zhang XS. Reconstruction of DNA sequencing by hybridization. Bioinformatics 2003; 19: 14-21.
38. Doi K, Imai H. Sequencing by hybridization in the presence of hybridization errors. Genome Inform Ser Workshop Genome Inform 2000; 11:53-62.
39. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, et al. Housekeeping genes as internal standards : use and limits. J Biotechnol 1999; 75:291-5.
40. Lee PD, Sladeck R, Greenwood CMT, Hudson TJ. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res 2002; 12:292-7.
41. Benes V, Muckenthaler M. Standardization of protocols in cDNA microarray analysis. Trends Biochem Sci 2003; 28: 244-9.
42. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, et al. An evaluation of the performance of cDNA microarray for detecting changes in global mRNA expression. Nucleic Acids Res 2001; 29:E41.
43. Badiee A, Eiken HG, Steen VM, Lovlie R. Evaluation of five different cDNA labelling methods for microarray using spike controls. BMC Biotechnol 2003; 3:23.
44. Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002; 32(Suppl):496-501.
45. Eickhoff B, Korn B, Schick M, Poustka A, van der Bosch J. Normalization of array hybridization experiments in differential gene expression analysis. Nucleic Acids Res 1999; 27:E33.
46. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 2001; 29:2549-57.
47. Yang IV, Chen E, Hasseman JP, Liang W, Wang S, Sharov V, et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol 2002; 3:research 0062.1-12.
48. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002; 30:E15.
49. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 2003; 4:210.
50. Nadon R, Shoemaker J. Statistical issues with microarrays: processing and analysis. Trends Genet 2002; 18:265-71.
51. Dudoit S, Gentleman RC, Quackenbush J. Open source software for the analysis of microarray data. BioTechniques 2003; 34(Suppl):45-51.
52. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 2002; 3:software0003.1-6.
53. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 2003; 34:374-8.
54. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5(10):R80.
55. Hardin J, Mitani A, Hicks L, VanKoten B. A robust measure of correlation between two genes on a microarray. BMC Bioinformatics 2007; 8:220.
56. Quackenbush J. Computational approaches to analysis of DNA microarray data. Methods Inf Med 2006; 45(Suppl 1):91-103.
57. Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 2003; 19:459-66.
58. Yin L, Huang CH, Ni J. Clustering of gene expression data: performance and similarity analysis. BMC Bioinformatics 2006; 7(Suppl 4):S19.
59. Hibbs MA, Dirksen NC, Li K, Troyanskaya OG. Visuali-

zation methods for statistical analysis of microarray clusters. BMC Bioinformatics 2005; 6:115.
60. Datta S, Datta S. Evaluation of clustering algorithms for gene expression data. BMC Bioinformatics 2006; 7(Suppl 4):S17.
61. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001; 2:418-27.
62. Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. Bioinformatics 2002; 18: 207-8.
63. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005; 4:Article 17.
64. Hartigan JA, Wong MA. A k-means clustering algorithm. Applied Statistics 1979; 28:100-8.
65. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceed Natl Acad Sci 1998; 95:14863-8.
66. Campagne F, Skrabanek L. Mining expressed sequence tags identifies cancer markers of clinical interest. BMC Bioinformatics 2006; 7:481-94.
67. Narayanan R. Bioinformatics approaches to cancer gene discovery. Methods Mol Biol 2007; 360:13-31.
68. Amatschek S, Koenig U, Auer H, Steinlein P, Pacher M, Gruenfelder A. Tissue-wide expression profiling using cDNA subtraction and microarrays to identify tumor-specific genes. Cancer Research 2004; 64:844-56.
69. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286:531-7.
70. RossDT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P,et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000; 24: 227-35.
71. Perou CM, Sorlie T, Eisen MB, van der Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumors. Nature 2000; 406:747-52.
72. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, et al. Genomic signatures to guide the use of chemotherapeutics. Nature Medicine 2006; 12:1294-1300.
73. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci 2001; 98:10869-74.
74.Ross js, Symmans WF, Pusztai L, Hortobagyi GN. Pharmacogenomics and clinical biomarkers in drug discovery and development. Am J Clin Pathol 2005; 124 (Suppl): S29-41.
75. Blower PE, Yang C, Fligner MA, Verducci JS, Yu L, Richman S, et al. Pharmacogenomic analysis: correlating molecular substructure classese with microarray gene expression data. Pharmacogenomics J 2002; 2:259-71.
76. Boyer J, Allen WL, McLean EG, Wilson PM, McCulla A, Moore S. Pharmacogenomic identification of novel determinants of response to chemotherapy in colon cancer. Cancer Res 2006; 66:2765-77.
77. Guerreiro N, Staedtler F, Grenet O, Kehren J, Chibout SD. Toxicogenomics in drug development. Toxicol Pathol 2003; 31:471-9.
78. Salter AH. Large-scale databases in toxicogenomics. Pharmacogenomics 2005; 6:749-54.
79. Newton RK, Aardema M, Aubrecht J. The utility of DNA microarrays for characterizing genotoxicity. Environ Health Perspect 2004; 112:420-2.
80. Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, et al. Development of large-scale chemogenomics database to improve selection and to understand mechanisms of chemical toxicity and action. J Biotechnol 2005; 119:219-44.
81. MAQC Consortium; Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006; 24:1151-61.
82. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol 2006; 24:1140-50.
83. Tong W, Lucas AB, Shippy R, Fan X, Fang H, Hong H, et al. Evaluation of external RNA controls for the assessment of microarray performance. Nat Biotechnol 2006; 24:1132-9.
84. Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, et al. Using RNA sample titrations to assess microarray platform performance and normalization techniques. Nat Biotechnol 2006; 24:1123-31.
85. Canales RD, Luo Y, Willey JC, Austermiller B, Barbaciouru CC, Boysen C, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol 2006; 24:1115-22.
86. Minor JM. Microarray quality control. Methods Enzymol 2006; 411:233-55.
87. Gulmann C, Sheehan KM, Kay EW, Lotta LA, Petricoin EF 3rd. Array-based proteomics: mapping of circuitries for diagnostics, prognostics, and therapy guidance in cancer. J Pathol 2006; 208:595-606.
88. Haab BB. Applications of antibody array platforms. Curr Opin Biotechnol 2006; 17(4):415-21.
89. Hiller R, Laffer S, Harwanegg C, Huber M, Schmidt WM, Twardosz A, et al. Microarrayed allergen molecules: diagnostic gatekeepers for allergy treatment. The FASEB Journal 2002; 16:414-6.
90. Kalbas M, Lueking A, Kowald A, Muellner S. New analytical tools for studying autoimmune diseases. Curr Pharm Des 2006; 12:3735-42.
91. Renault NK, Mirotti L, Alcocer MJ. Biotechnologies in new high-throughput food allergy tests: why we need them. Biotechnol Lett 2007; 29:333-9.
92. Harwanegg C, Hiller R. Protein microarrays in diagnosing IgE-mediated diseases: spotting allergy at the molecular level. Expert Rev Mol Diagn 2004; 4:539-48.
93. Casiano CA, Mediavilla-Varela M, Tan EM. Tumor-associated antigen arrays for the serological diagnosis of cancer. Mol Cell Proteomics 2006; 5:1745-59.
94. Draghici S, Chatterjee M, Tainsky MA. Epitomics: serum screening for the early detection of cancer on microarrays using complex panels of tumor antigens. Expert Rev Mol Diagn 2005; 5:735-43.
95. Caron M, Choquet-Kastylevsky G, Joubert-Caron R. Cancer immunomics using autoantibody signatures for biomarker discovery. Mol Cell Proteomics 2007; 6:1115-22.
96. Robinson WH. Antigen arrays for antibody profiling. Curr

Opin Chem Biol 2006; 10:67-72.
97. Hueber W, Kidd BA, Tomooka BH, Lee BJ, Bruce B, Fries JF, et al. Antigen microarray profiling of autoantibodies in rheumatoid arthritis. Arthritis Rheum 2005; 52: 2645-55.
98. Li QZ, Zhou J, Wandstrat AE, Carr-Johnson F, Branch V, Karp DR, et al. Protein array autoantibody profiles for insights into systemic lupus erythematosus and incomplete lupus syndromes. Clin Exp Immunol 2007; 147:60-70.
99. Sartain MJ, Slayden RA, Singh KK, Laal S, Belisle JT. Disease state differentiation and identification of tuberculosis biomarkers via native antigen array profiling. Mol Cell Proteomics 2006; 5:2102-13.
100. Binder SR, Hixson C, Glossenger J. Protein arrays and pattern recognition: new tools to assist in the identification and management of autoimmune disease. Autoimmun Rev 2006; 5:234-41.
101. Haab BB. Antibody arrays in cancer research. Mol Cell Proteomics 2005; 4:377-83.
102. Sanchez-Carbayo M. Antibody arrays: technical consideration and clinical applications in cancer. Clin Chem 2006; 52(9):1651-9.
103. Belov L, Mulligan SP, Barber N, Woolfson A, Scott M, Stoner K, et al. Analysis of human leukaemias and lymphomas using estensive immunophenotypes from an antibody microarray. Br J Haematol 2006; 135:184-97.