TdG e microarray

Collage a cura di
Fioravante PATRONE
http://www.diptem.unige.it/patrone/default.htm

Si tratta di un collage di documenti già disponibili in rete:

- appunti di Moretti per il corso di Teoria delle Decisioni per SMID

- una pag. cadauno dalle relazioni di Bonassi, Varesio e Moretti al Festival della Scienza 2007

Fioravante PATRONE
Dipartimento di Ingegneria della Produzione, Termoenergetica e Modelli Matematici
P.le Kennedy - Pad D
16129 Genova - ITALY
patrone@diptem.unige.it

http://www.diptem.unige.it/patrone	homepage
http://tdg.dima.unige.it	web teaching
http://www.citg.unige.it/citg.htm	web server "CITG"
http://www.scallywag.it	web page del gruppo Scaλλywag
http://www.diptem.unige.it/patrone/DRI.htm	Decisori (razionali) interagenti

Teoria dei Giochi applicata all'analisi di espressione genica

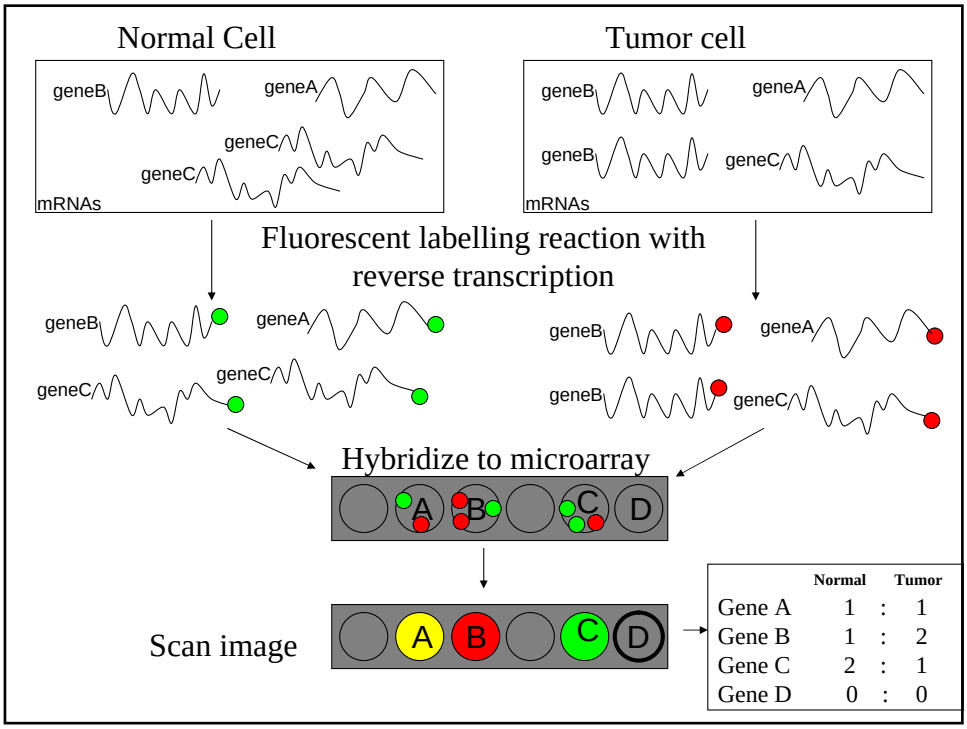
Stefano Moretti – Corso di Teoria delle Decisioni

Anno 2006

Lezione 23/11/06

Espressione genica e microarray

- *L' espressione genica* quando l'informazione genetica contenuta nel DNA è *transcritta* in molecole di mRNA e quindi *tradotta* in proteine.
- Stime di quantità mRNA forniscono una valutazione di quante proteine siano codificate da un dato gene, ovvero la sua espressione genica.
- La tecnologia basata sui microarray permette di scattare fotografie dell'espressione genica. All'interno di un singolo esperimento è possibile stimare il livello di espressione di migliaia di geni sotto la stessa condizione biologica (per es. Tumore)


Normal Cell
Tumor cell
geneB
geneA
geneC
geneC
mRNAs
geneB
geneA
geneB
geneC
mRNAs
Fluorescent labelling reaction with
reverse transcription
geneB
geneA
geneC
geneC
geneB
geneA
geneB
geneC
Hybridize to microarray
A
B
C
D
Scan image
A
B
C
D
Normal Tumor
Gene A 1 : 1
Gene B 1 : 2
Gene C 2 : 1
Gene D 0 : 0

...

Array1 Array2 Array3

Arrays

Genes	Array1	Array2	Array3	Array4	Array5	...
Gene1	0.46	0.30	0.80	1.51	0.90	...
Gene2	0.10	0.49	0.24	0.06	0.46	...
Gene3	0.15	0.74	0.04	0.10	0.20	...
Gene4	0.45	1.03	0.79	0.56	0.32	...
Gene5	0.06	1.06	1.35	1.09	1.09	...
...	...	...	...	...	...	

Livello di espressione del gene 5 nel campione 4

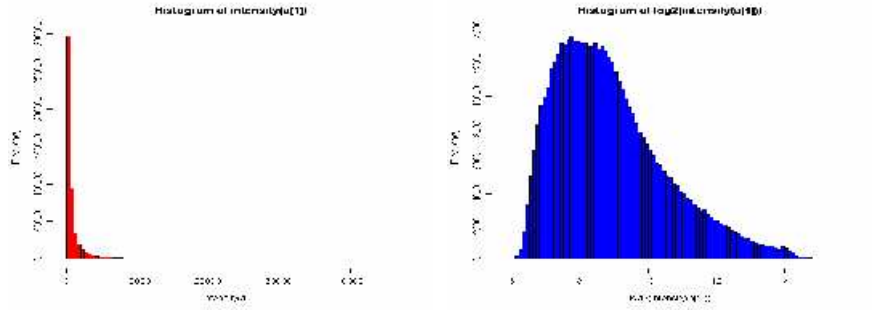
Matrix:

Arrays

Genes	Array1	Array2	Array3	Array4	Array5	...
Gene1	0.46	0.30	0.80	1.51	0.90	...
Gene2	-0.10	0.49	0.24	0.06	0.46	...
Gene3	0.15	0.74	0.04	0.10	0.20	...
Gene4	-0.45	-1.03	-0.79	-0.56	-0.32	...
Gene5	-0.06	1.06	1.35	1.09	-1.09	...
...	...	...	...	...	...	

M = $\log_2$(Red intensity / Green intensity)

Logaritmo del valore di espressione



- Migliora la descrizione dei dati: riduce la “skewness” della distribuzione rendendo più simmetrica la distribuzione dell’intensità dei probe
- La variazioni del logaritmo delle intensità tendono ad essere meno dipendenti dalla magnitudo del segnale; come conseguenza viene meglio stimata anche la varianza

Sorgenti di variazione dell’espressione genica

- Alcune variazioni osservate sono dovute a risposta diverse a condizioni genetiche e ambientali differenti (es. cellule malate e cellule sane): *variazione interessante*.
- Altre variazioni sono introdotte per errore durante
  - la preparazione dei campioni
  - La realizzazione degli array
  - Il processamento degli array (labeling, ibridizzazione, scannerizzazione)
- trattasi di *variazione oscura* che deve essere eliminata attraverso il processo di ***normalizzazione***

Normalizzazione

"Soluzione": trovare un insieme di *geni invarianti* cioè tali che

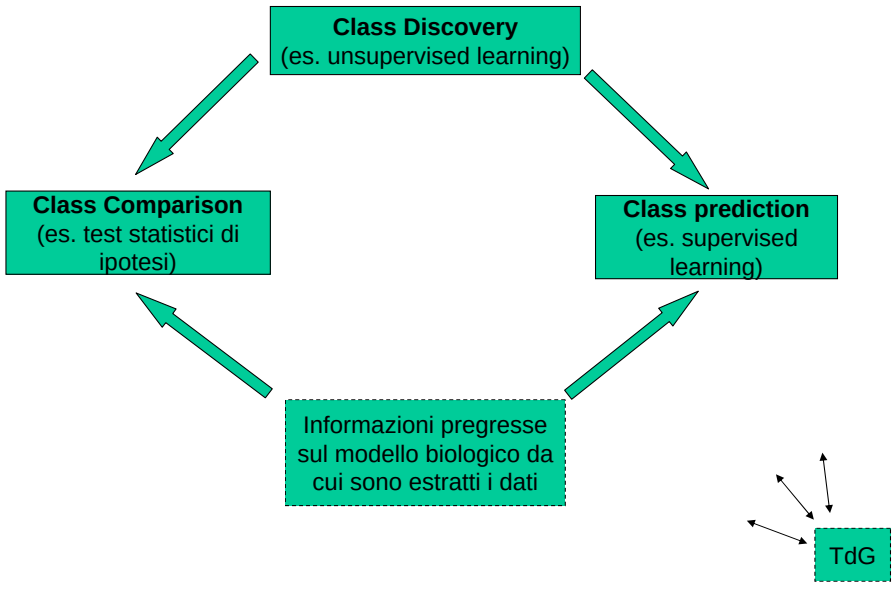
1) i loro valori di espressione rimangano costanti su tutti gli array

2) i loro valori di espressioni ricoprano l'intero spettro di intensità del segnale osservato. (NB: Il fattore di normalizzazione necessario per aggiustare le intensità basse non necessariamente è uguale a quello utilizzato ad intensità elevate).

3) i rapporti di normalizzazione tra questi geni siano rappresentativi dei rapporti di normalizzazione per tutti i geni

- *Geni di controllo* (spiked-in): geni sintetici a concentrazioni note (3?)
- *Geni housekeeping*: geni che sono assunti (in partenza) essere invarianti tra array differenti (1? e 2?)
- *Geni Invarianti*: geni che vengono osservati, secondo qualche metrica, come poco variabili lungo gli array
- *Tutti i geni*: è ragioveole aspettarsi che siano molto pochi i geni che variano a causa di una diversa risposta a condizioni di interesse differenti (più è piccolo il numero, più 1 è soddisfatto)

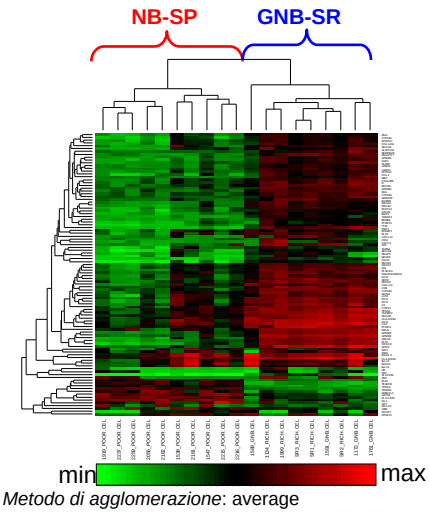
Analisi statistica di microarray



Class Discovery

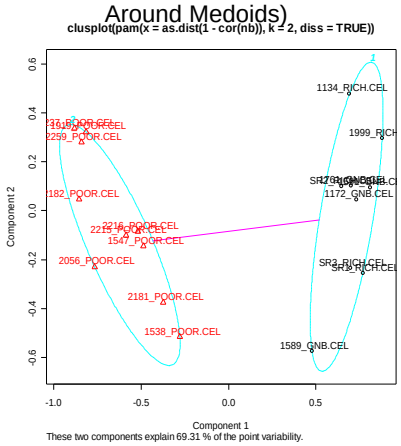
- Preselezione rozza (non necessaria): Primi 100 geni con la più alta deviazione standard.
- Risultato sul dataset con 19 campioni:

Hierarchical Clustering



Metodo di agglomerazione: average linkage; *Metrica*: euclidea

PAM (Partitioning Around Medoids)



Algoritmo: simile al K-means

Numero di classi in input:2

Metrica: 1-correlazione

Class comparison

- Nei test statistici di ipotesi la congettura che non ci sia differenza tra due classi è detta *ipotesi nulla*
- Il risultato di un test di ipotesi è una *decisione:*
  - *Rifiutare* l'ipotesi nulla e asserire che c'è una differenza tra le due classi (risultato *positivo*)
  - *Non rifiutare* l'ipotesi nulla e dichiarare che c'è un'evidenza insufficiente per determinare una differenza tra le due classi (risultato *negativo*)
- Se la decisione del test è di rifiutare l'ipotesi nulla, può succedere che
  - si stia rifiutando correttamente una ipotesi nulla che è falsa (risultato *vero positivo*)
  - si stia rifiutando scorrettamente una ipotesi nulla che è vera (risultato *falso positivo*)
- Il livello del test è la probabilità di avere un risultato falso positivo e si indica generalmente con α

Test multipli

- Con i dati di microarray si ha a che fare con G ipotesi nulle da testare, dove G è il numero di geni analizzati; la i-esima ipotesi nulla, per i=1,…,G, esprime la congettura che il gene i non sia differenzialmente espresso tra le due classi
- Se ogni test è applicato ad un livello $\alpha \in [0,1]$ (e i test sono indipendenti) la probabilità p di avere almeno un falso positivo è $1-(1-\alpha)^G$.
- Inoltre il numero atteso di falsi r positivi è αG.
- Per controllare p ed r si può fissare un livello α molto piccolo
- E allora rischierei di non avere geni significativi (leggi differenzialmente espressi)

Class prediction

- A partire dai dati di cui si conosce la classe di provenienza (*training set*) si ricava una "*regola*" basata sui valori di espressione di un sottogruppo di geni in grado di collocare correttamente i campioni nelle due classi
- La speranza è che tale regola sia anche in grado di collocare correttamente un campione la cui origine è sconosciuta (quindi apre possibilità per diagnosi precoce, regole per la decisione della terapia ecc.)
- Bioconductor: libreria *pamr* software per la classificazione basato sul metodo "Diagnosis of multiple cancer types by shrunken centroids of gene expression" Tibshirani et al. PNAS (2002) vol. 99 no. 10 **6567–6572**

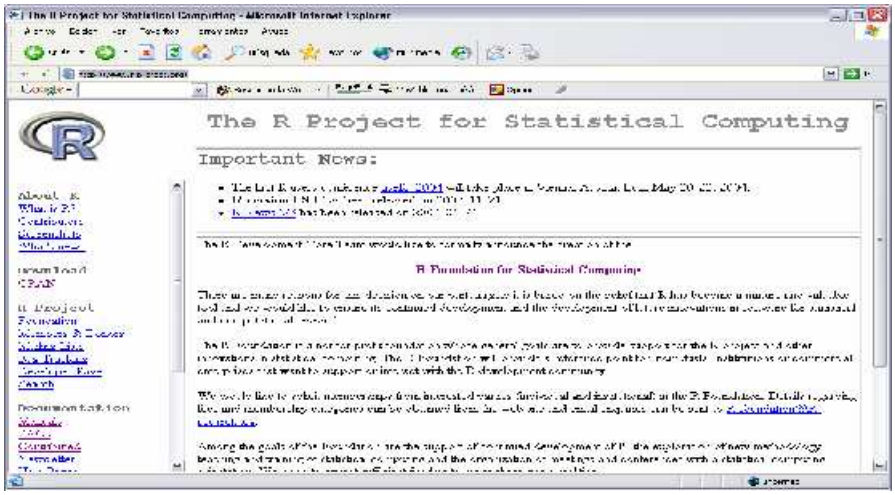
Alcuni metodi usati in class prediction

- Metodi statistici classici parametrici
  - Metodo della massima verosomiglianza
  - Metodo dello stimatore bayesiano
- Metodi statistici classici non parametrici
  - Parzen windows
  - K_m nearest neighbor
  - Nearest neighbor
  - Separatori lineari (es. discrim. Di Fischer, percettrone)
  - Metodi euristici (es. Case-based reasoning)
- Metodi basati su espansione in serie
  - Reti neurali
  - Radial basis function
  - Support vector machine
- Alberi decisionali

R

- R è un sistema open-source per l'analisi statistica e la descrizione dei dati. Consiste di
  - Un linguaggio
  - Un ambiente run-time con
    - Finestra grafici, debugger, ecc
    - Accesso ad alcuni sistemi di funcioni,
- Può essere usato come
  - Interattivo, attraverso un linguaggio di comando
  - O esecuzione di programmi immagazzinati in file script

http://www.r-project.org/



Il progetto Bioconductor

- Progetto *Open source* e *open development software* per l'analisi e la comprensione di dati genomici.
- La maggior parte dei programmi è stata pensata come librerie di R.
- Documentazione esaustiva e materiale didattico al sito http://www.bioconductor.org/
- Ha raggiunto una certa stabilità ma si sta ancora evolvendo. → *quello che ora sembra uno standard potrebbe non esserlo più nel prossimo futuro.*
- *Ha raggiunto la versione 2.2*

Questione aperta

In accordo ai dati di microarray di espressione raccolti, è possibile quantificare il "potere" dei geni nel determinare l'insorgenza della malattia genetica di interesse?

Expression data for a three genes microarray experiment on three tumor cells

	t1	t2	t3
g1	**0.2**	**20**	**12**
g2	**11**	**9.8**	**8.6**
g3	**-7**	**-0.1**	**-6.1**

Discretized matrix

	t1	t2	t3
g1	0	1	1
g2	1	1	1
g3	1	0	1

- **M≅0**, the gene is *normally expressed* in the tumor cell → *0*
- **M>>0** o **M<<0**, the gene is *abnormally expressed* in the tumor cell → *1*

Discretized matrix

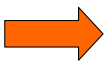
	s1	s2	s3
g1	0	1	1
g2	1	1	1
g3	1	0	1

In accordo ai dati di microarray di espressione raccolti, è possibile quantificare il "potere" dei geni nel determinare l'insorgenza della malattia genetica di interesse?

Scendono in campo i geni …

- I giocatori sono proprio i geni
- Ma chi fornisce la **regola decisionale** nel contesto dei geni?
- Possibile risposta: affidiamoci ai dati di espressione forniti dai microarray.
- Esempio: definiamo un criterio per stabilire quali geni si comportano in maniera "anormale" su ciascun array.

	array1
gene1	0.121
gene2	2.453
gene3	3.586



	array1
gene1	0
gene2	1
gene3	1

Regola decisionale

Un gruppo di geni è "*vincente*" in un array se tutti i geni che si comportano in maniera "anormale" nell' array sono contenuti nel gruppo.

	t1
gene1	0
gene2	1
gene3	1

Sia gruppo {gene2, gene3} che il gruppo {gene1, gene2, gene3} è vincente.

Microarray game:

- Players are genes;
- games with [0,1]-characteristic function;
- (*sufficiency principle*) on each sample:
  - a coalition S⊆N is a *winning coalition in a sample j* if **all** the abnormally expressed genes in the sample *j* belong to S;
  - otherwise, a coalition S⊆N is a *loosing coalition in a sample j*.

EXAMPLE:

Microarray discr. data

	s1	s2	s3
g1	0	1	1
g2	1	1	1
g3	1	0	1

The corresponding *microarray game* <{g1,g2,g3},v> is:

v({g1,g2})=v({g3,g2})=1/3
v({g1,g2,g3})=1 and
v(S)=0 for each other coalition S.

The Shapley value is: (5/18,8/18,5/18).

Q: *Which properties a power index should satisfy for fairly measuring the power of genes in activating the disease?*

Property 1: null player (NP)

A gene which does not contribute to change the value (of activations of the tumor) of any coalition of genes, should receive zero power.

Prop. 2:equal splitting (ES)

Each sample should receive the same level of reliability. So the power of a gene on two samples should be equal to the sum of the power on each sample divided by two.

	s1		s2			s1	s2	
g1	0	ψ_1	1	ψ'_1	g1	0	1	$(\psi_1+\psi'_1)/2$
g2	0	ψ_2	1	ψ'_2	g2	0	1	$(\psi_2+\psi'_2)/2$
g3	1	ψ_3	0	ψ'_3	g3	1	0	$(\psi_3+\psi'_3)/2$

Partnership of genes

A group of genes S such that does not exist a proper ($\subset$) subset of S which contributes in changing the number of activations of the disease of groups of genes outside S.

Example

These two sets are partnerships of genes in the corresponding Microarray game

	s1	s2	s3
g1	0	1	1
g2	0	1	1
g3	1	0	1

Property 3: partnership monotonicity (PM)

Consider two disjoint partnerships of genes S,T with the same value and such that $v(S \cup T)=v(N)$. Then genes in the smaller one should receive not less power index than genes in the larger one.

Example

		s1	s2
ψ_1	g1	0	1
ψ_2	g2	0	1
ψ_3	g3	1	0
ψ_4	g4	1	0
ψ_5	g5	1	0

$\psi_i = \psi_k$

For each $i \in \{1,2\}$ $k \in \{3,4,5\}$

Property 4: partnership rationality (PR)

The total amount of power index received from players of a partnership S should not be less than v(S)

Property 5: partnership feasibility (PF)

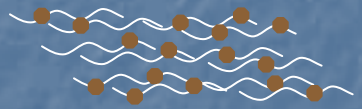
The total amount of power index received from players of a partnership S should not be greater than v(N)

***Theorem*:** The Shapley value is the unique solution which satisfies NP, ES, PM, PR, PF on the class of Microarray games.

GeneChip Experimental Workflow

Isolate total RNA from cells

Enzymatic amplification to generate biotin-labeled cRNA (50-100 fold amplification)

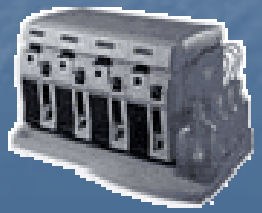


Hybridize to Array (45°C overnight)

#	UniqueID	Function	Accession	Clone	Ratio	Ratio 1	Ratio 2	Ratio 3
1	380	(D00710) heat-shock protein [	MEN	4h08f1	6.55	7.32	7.75	5.20
2	6651	heat shock protein 101 (HSP101	plW24	plW24	6.19	7.46	4.70	6.32
3	5168	(Y11828) heat shock protein [	MEN	58g08f1	5.79	6.82	6.57	4.35
4	1258	HEAT SHOCK PROTEIN 81-2 (HSP8	MEN	14a10f1	5.54	5.20	5.63	5.66
5	5309	BETA-AMYLASE (1 4-ALPHA-D-GL	MEN	31c05f1	5.47	6.25	4.10	6.36
6	2280	hypothetical protein T9E8.150	MEN	25f12f1	5.29	7.07	4.33	4.78
7	6267	heat shock protein 101 (HSP101	plW24	plW24	5.15	5.65	4.99	4.90
8	6290	HSP70-3cyt	HSP70-3cyt	slW217	5.08	5.53	4.20	4.42
9	5584	(AC003105) unknown protein [A	MEN	73b04f1	4.87	4.74	4.06	5.53
10	1234	(AC002337) putative galactino	MEN	13g10f1	4.57	5.15	3.99	3.88

Data Extraction

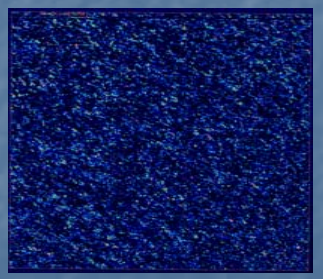
Wash & Stain



Fluidics

Image Capture

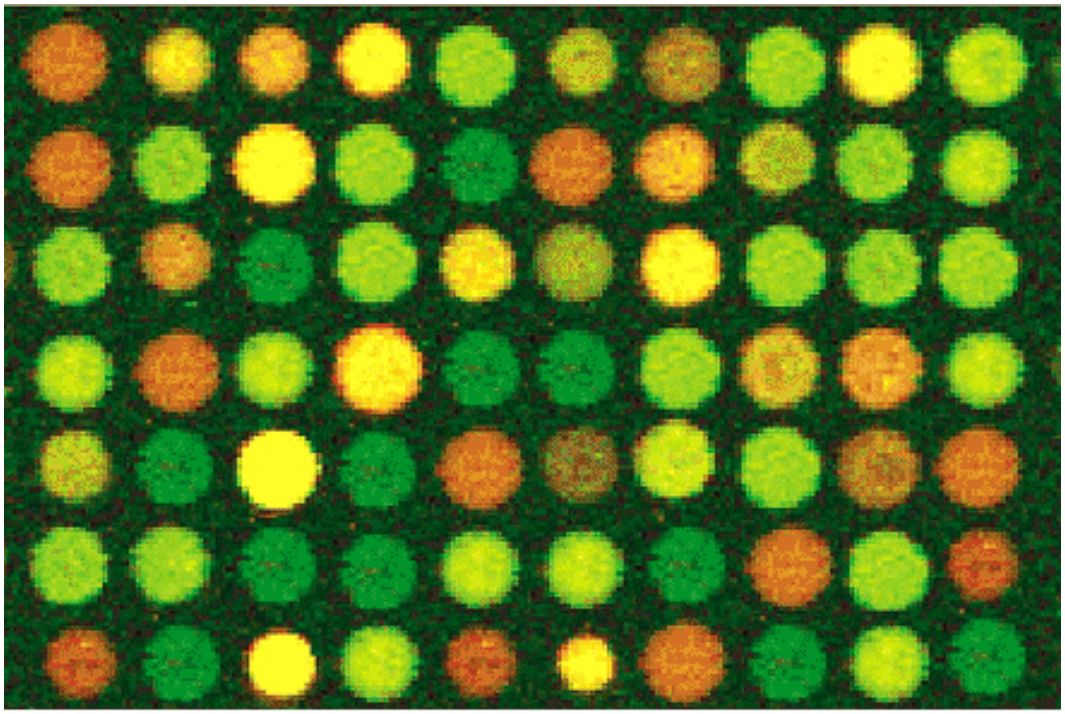
Scanner



16 bit TIFF image

Microarray to be used as routine clinical screen

C. M. Schubert
Nature Medicine
*9, **9,** 2003.*



The Netherlands Cancer Institute in Amsterdam is to become the first institution in the world to use microarray techniques for the routine prognostic screening of cancer patients. Aiming for a June 2003 start date, the center will use a panoply of 70 genes to assess the tumor profile of breast cancer patients and to determine which women will receive adjuvant treatment after surgery.

Applicazioni

- In collaborazione con l'Unità di Pediatria Oncologica Translazionale abbiamo analizzato *22283* sequenze genomiche da cellule di neuroblastoma, tumore maligno embrionario specifico del bambino.
- Sono stati selezionati i 50 geni più importanti in base all'indice di potere di Shapley & Shubik.
- Tra questi era presente il gene *MYCN,* già noto dalla letteratura per la sua associazione con il neuroblastoma.
- Stiamo studiando i rimanenti 49…