Riconoscimento e recupero dell’informazione per bioinformatica

Analisi di dati di espressione

Manuele Bicego

Corso di Laurea in Bioinformatica
Dipartimento di Informatica - Università di Verona

Sommario

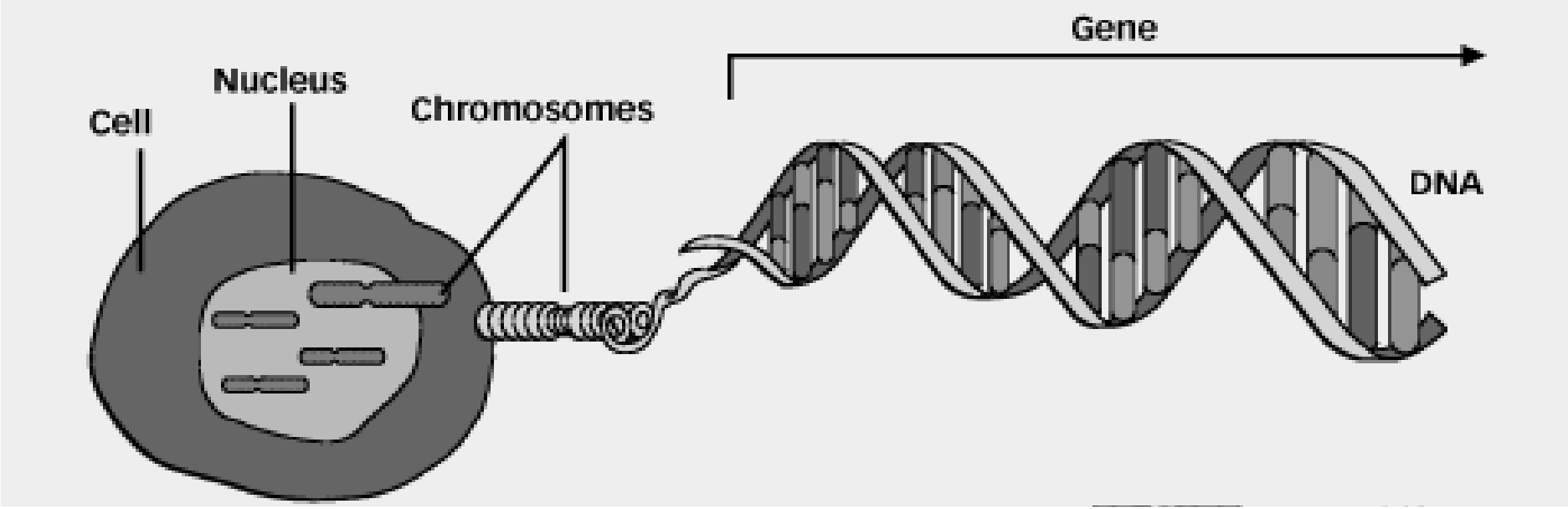
- Introduzione all'espressione genica
  - cos’è, come funziona, qual'è l'output

- Problematiche di image processing (per Microarray)
- Analisi dei dati di espressione genica
  - Analisi statistica
  - Classificazione di esperimenti
  - Clustering di dati di espressione

Analisi di dati espressione con tecniche di modellazione linguistica

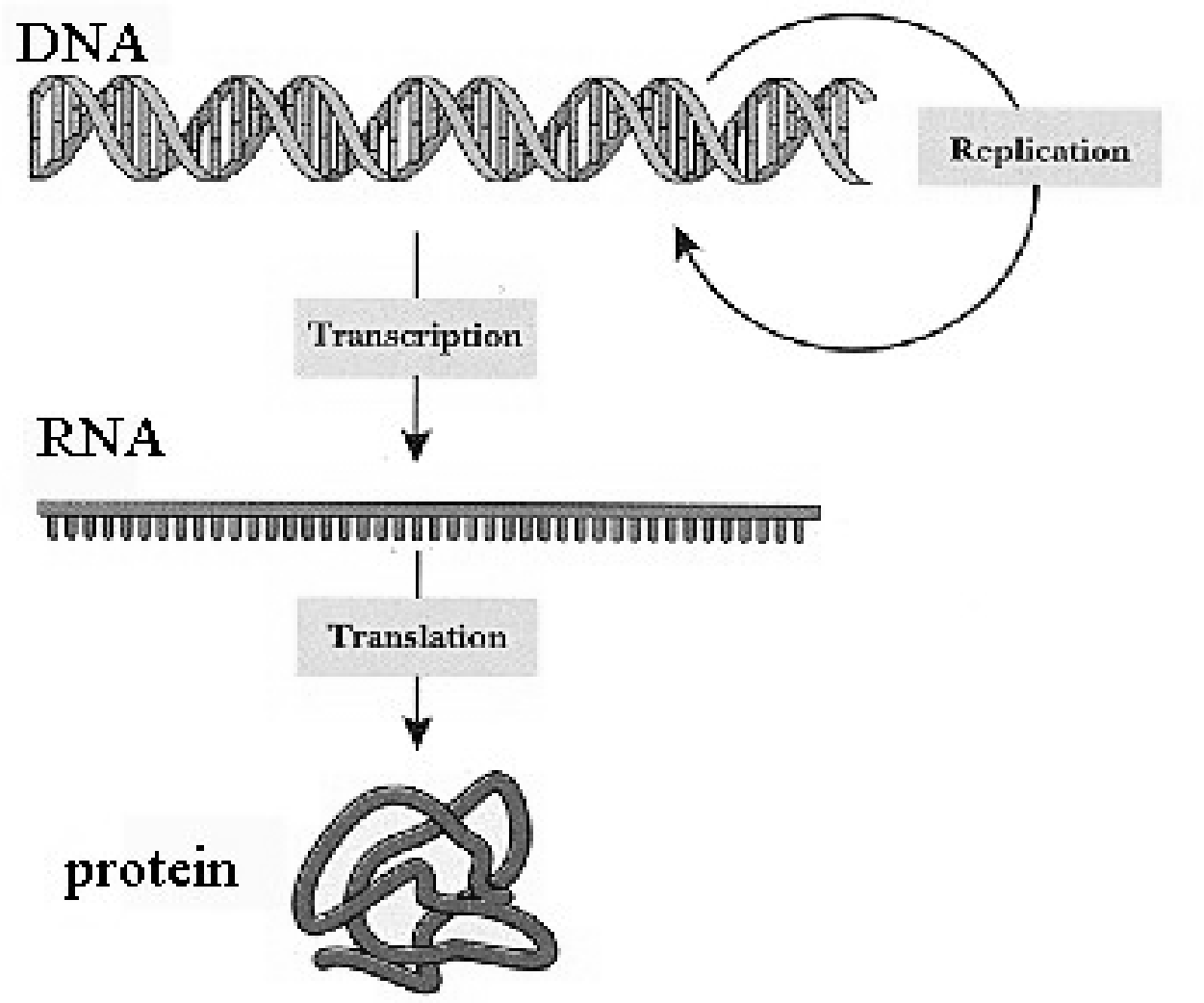
Espressione genica

Background



DNA: informazione genetica di una cellula

Dal DNA alla funzione



Step 1: Trascrizione: durante la trascrizione il DNA viene trascritto nel mRNA (messenger ribonucleid acid)

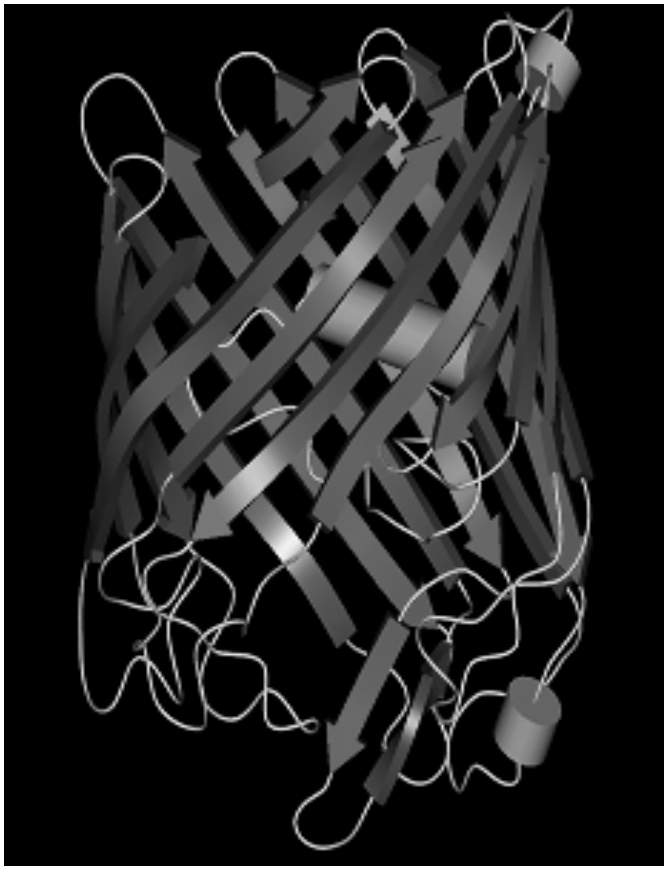
Step 2: Traduzione: durante la traduzione l-mRNA viene tradotto per produrre una proteina

Background: formato dei dati

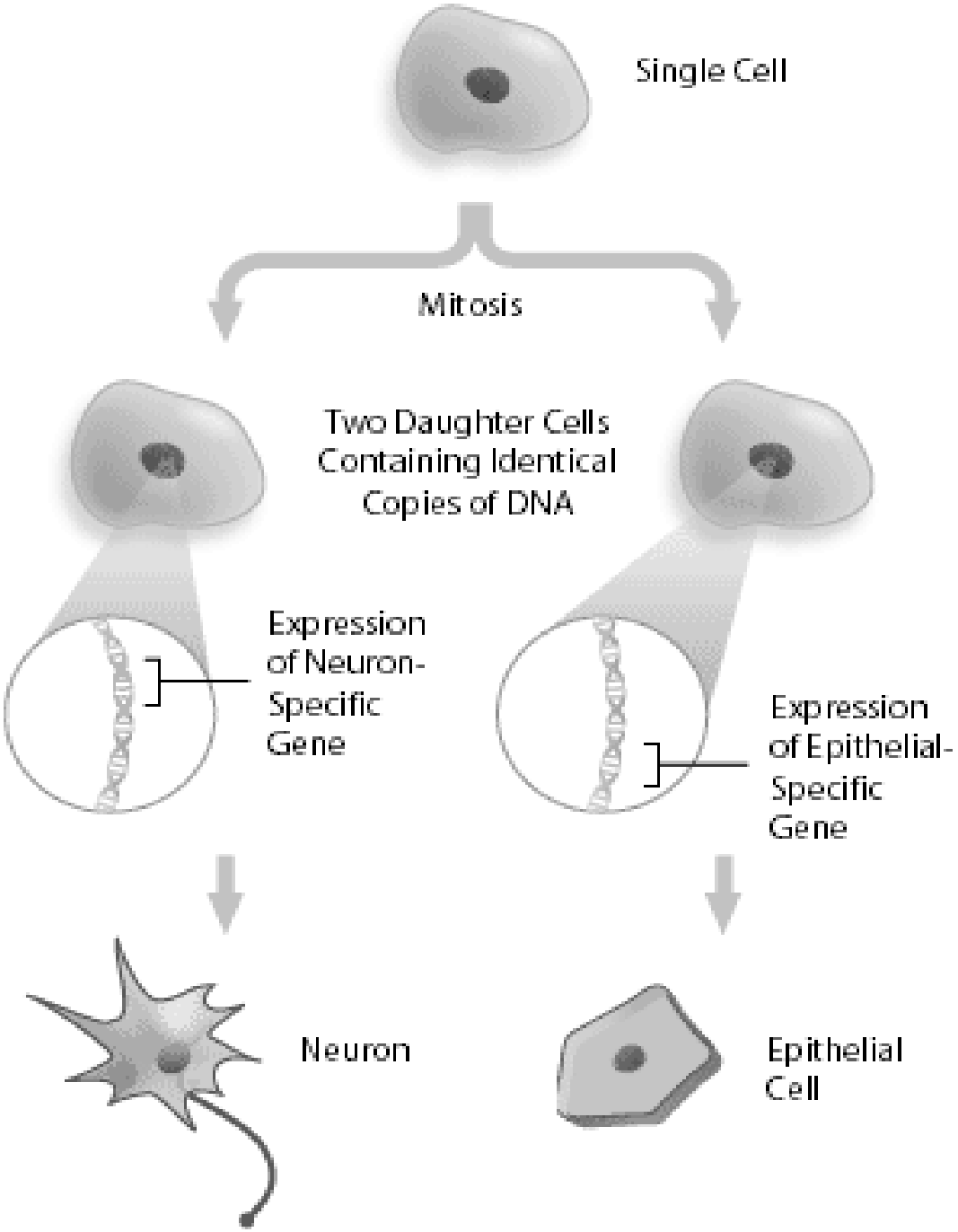
```
atgcgat
cgatcga
tcgatca
ggcgcgc
tacgagc
ggcgagg
acctcat
catcgat
cag
```

```
augcgau
cgaucga
ucgauca
ggcgcgc
uacgagc
ggcgagg
accucau
caucgau
cag
```

```
MRPQAPGSLVDPNE
DELRMAPWYWGRIS
REEAKSILHGKPDG
SFLVRDALSMKGEY
TLTLMKDGCEKLIK
ICHMDRKYGFIETD
LFNSVVEMINYYKE
NSLSMYNKTLDITL
SNPIVRAREDEESQ
PHGDLCLLSNEFIR
TCQLLQNLEQNLEN
KRNSFNAIREELQE
KKLHQSVFGNTEKI
FRNQIKLNESFMKA
PADA……
```



DNA: informazione genetica → mRNA: informazione espressa → Sequenza proteica: informazione codificata → Proteina ripiegata: funzione



Osservazioni generali

- Il genoma non e' composto da soli geni
- Tutte le cellule di un organismo hanno lo stesso genoma
- Il tipo di cellula dipende dai geni "espressi"
  - Solo una frazione dei geni viene espressa
  - l'espressione non e' binaria
  - L'espressione di un gene dipende anche dalle condizioni ambientali

Espressione genica

- Il DNA determina il differente tipo di cellule utilizzando un meccanismo che si chiama “differential gene expression”
  - Questo meccanismo determina dove, quando e in quale quantita’ un gene e’ espresso in una cellula in un determinato momento
  - Questo processo permette di produrre differenti tipi di cellule utilizzando lo stesso genoma

Espressione genica

- Capire l’espressione dei geni ci permette di spiegare le funzioni della cellula ed eventualmente la sua patologia
- L’espressione di un gene e’ misurabile in termini di “abbondanza” di mRNA
- Queste misure possono essere effettuate con diverse tecniche (es. Microarray, RNAseq...)
- Si puo’ dire che queste tecnologie permettono di misurare il *livello di espressione* dei geni

I microarray

Tecnica estremamente utilizzata per misurare l'espressione genica (oggigiorno esistono alternative migliori – *next-generation sequencing (*RNA-seq))

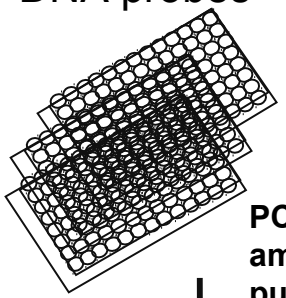
Punto di vista PR: ha senso studiarli, molti dati disponibili!

Terminologia:

- ⇨ array: substrato dove vengono immobilizzati i probes
- ⇨ I probes rappresentano i pezzi di DNA immobilizzati sull'array – i.e. il substrato immobile
- ⇨ Ibridazione: processo con cui si calcola l'espressione
- ⇨ I targets rappresentano le sequenze di cDNA che vengono ibridate sull'array – i.e. il substrato mobile

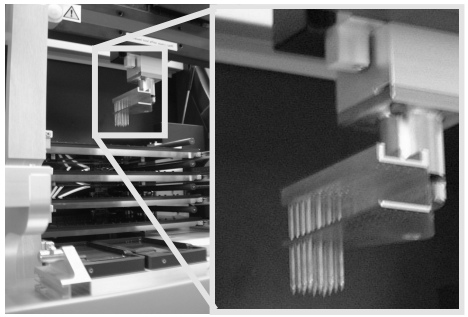
Procedura: due condizioni

DNA probes



PCR product amplification purification

printing



- Extract mRNA from samples.
- Reverse transcription of mRNA to cDNA.
- Label with Cy3 and Cy5 fluorescent dyes.
- Hybridize labeled cDNA cocktail to array.

Test sample

Reference sample

RNA

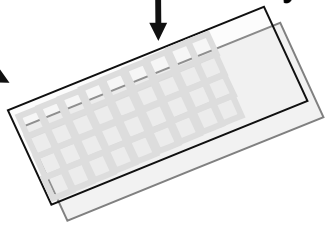
RNA

cDNA

cDNA

Hybridize

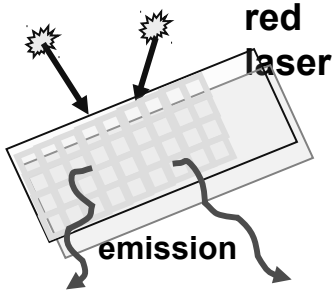
Production



excitation

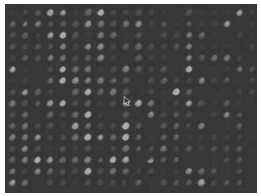
green laser

red laser

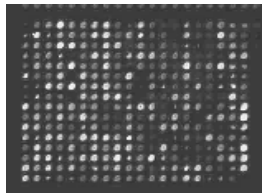


emission

scanning

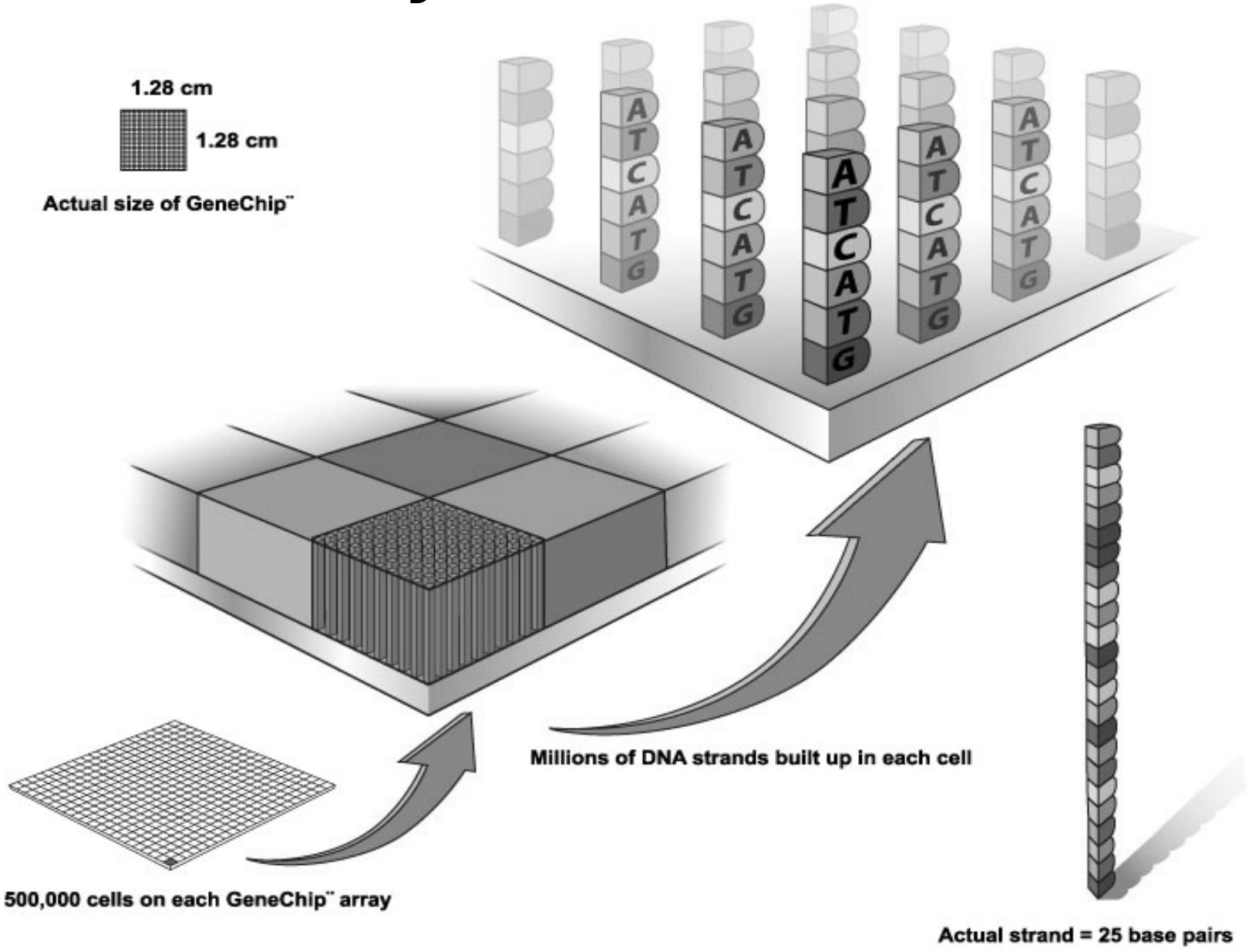


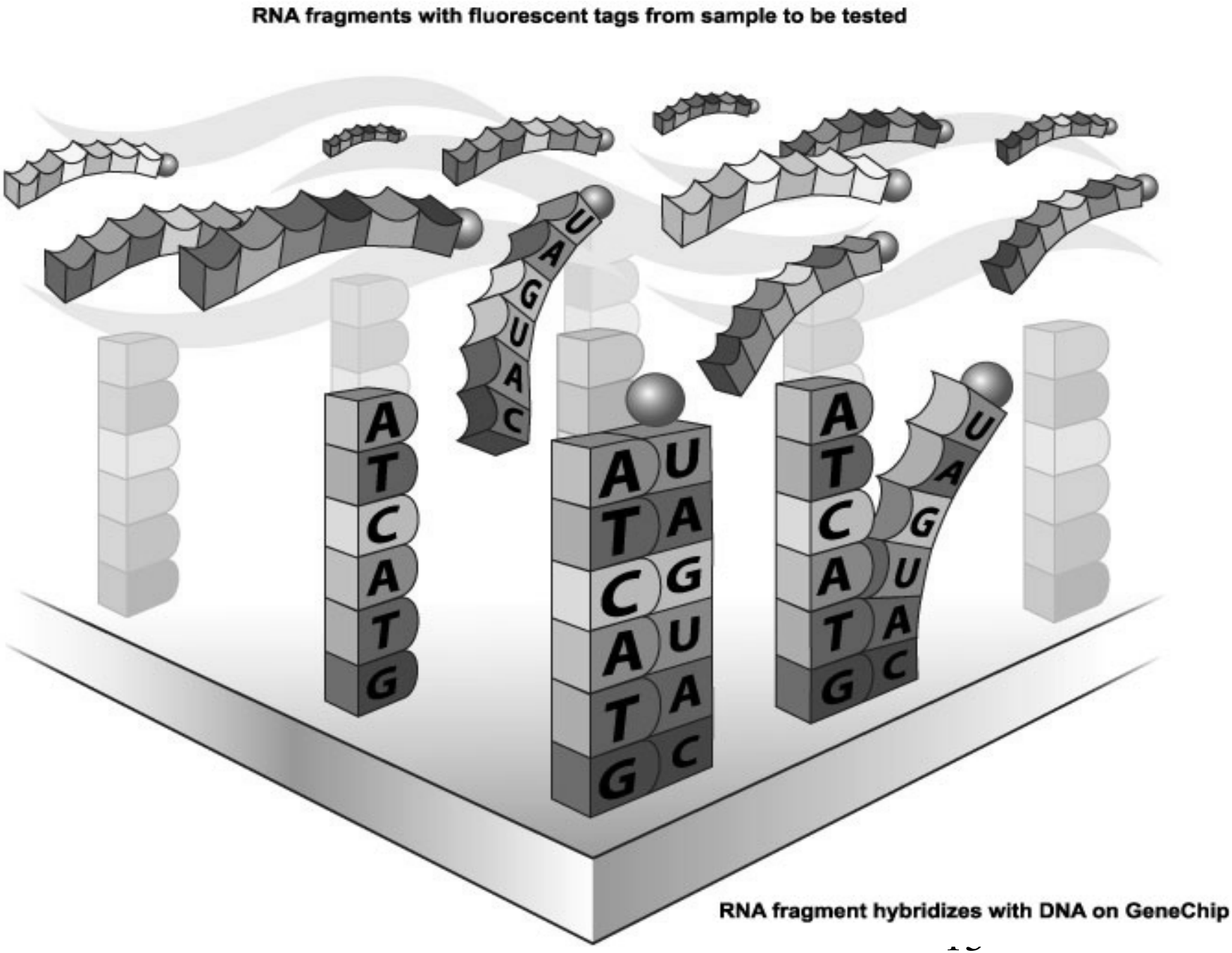
overlay images



data: (R_{fg},G_{fg},R_{bg},G_{bg}, ...)

Microarray: il meccanismo




RNA fragments with fluorescent tags from sample to be tested
RNA fragment hybridizes with DNA on GeneChip

RNA Seq

RNA seq: Next Generation DNA sequencing

- ⇨ Tecnologia in grado di contare accuratamente I trascritti e comparare i campioni (può generare più di un milione di RNA sequences per campione!)
- ⇨ “Digital Gene Expression”
- ⇨ Può anche identificare isomorfismi, varianti etc etc

Analisi di dati di espressione genica

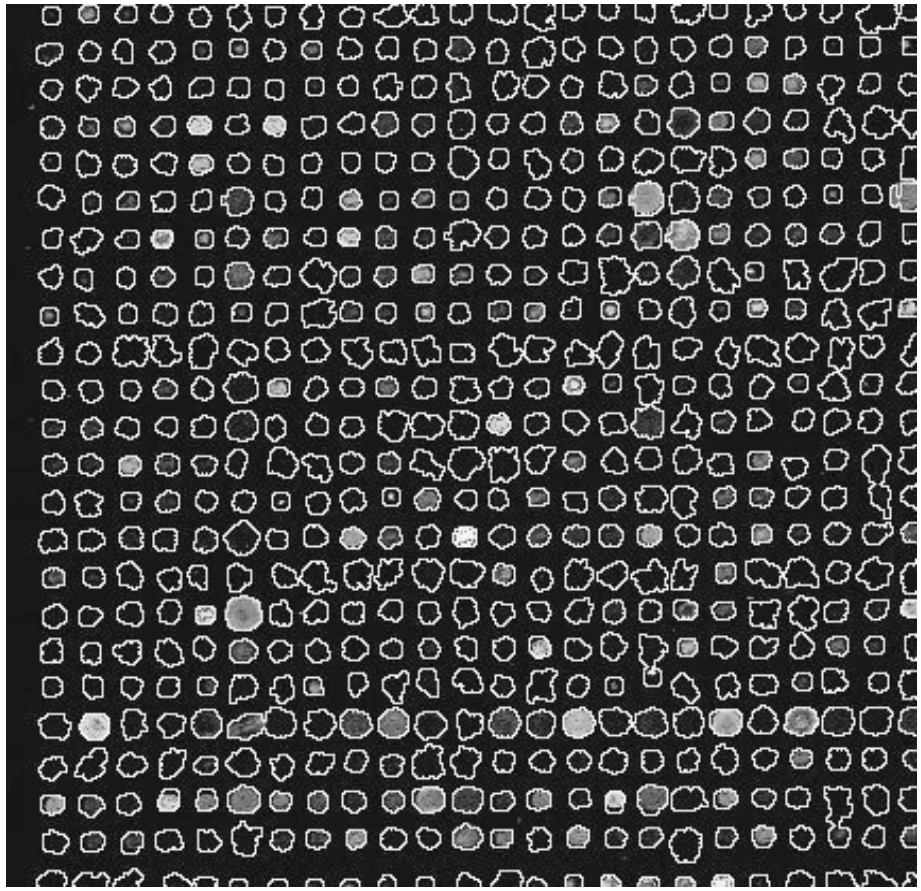
Problematiche

- Problematiche di image processing (per i microarray)
  - Segmentazione spots e rimozione rumore
  - Quantificazione del segnale
  - Rilevamento della qualita’
- Problematiche di pattern recognition
  - Analisi statistica
  - Classificazione di esperimenti
  - Clustering di geni/esperimenti e Bi-clustering

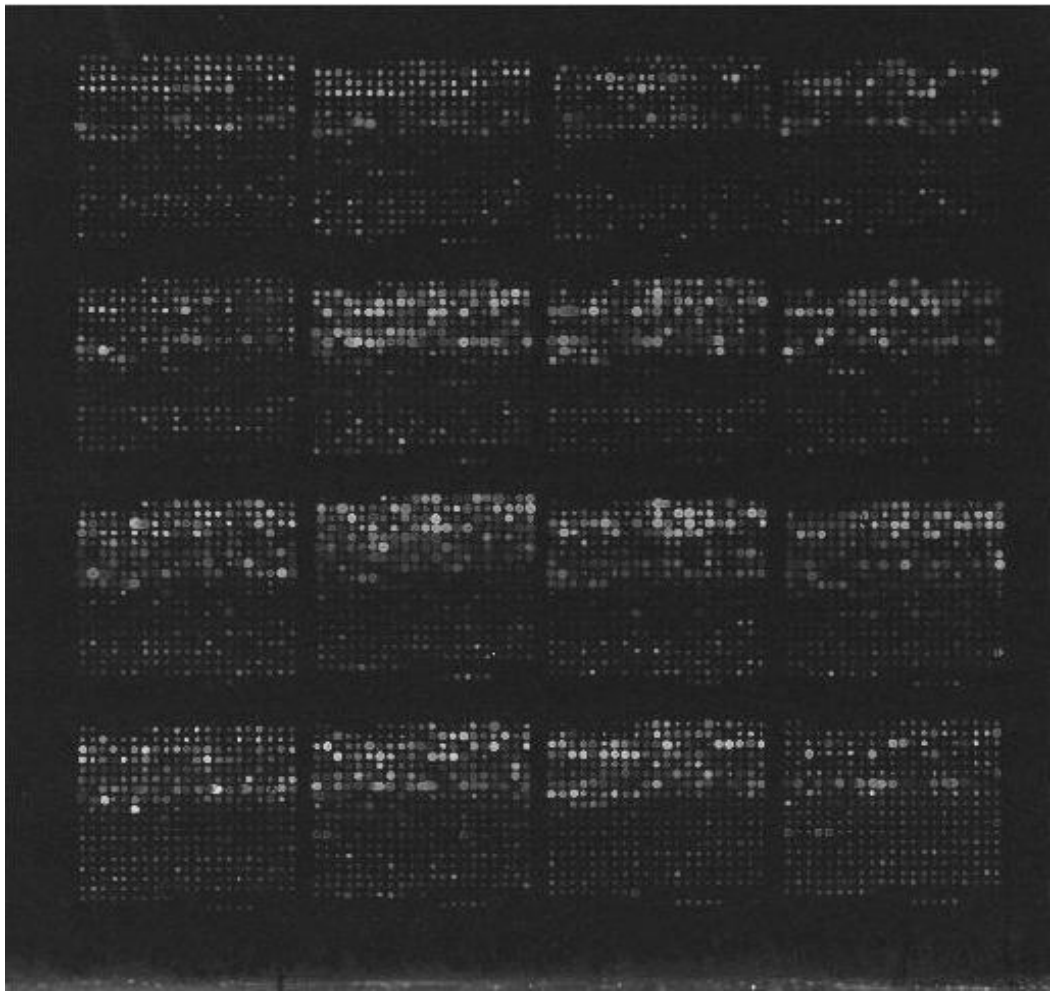
Problemi di Image processing

1. Identificare gli spot
2. Preprocessing (normalizzazione) e segmentazione
   - Classificazione di pixel come segnale o background
3. Quantificazione del segnale

   a) stima del foreground
   b) stima del background
   c) ... (shape, size etc)
5. Qualita’

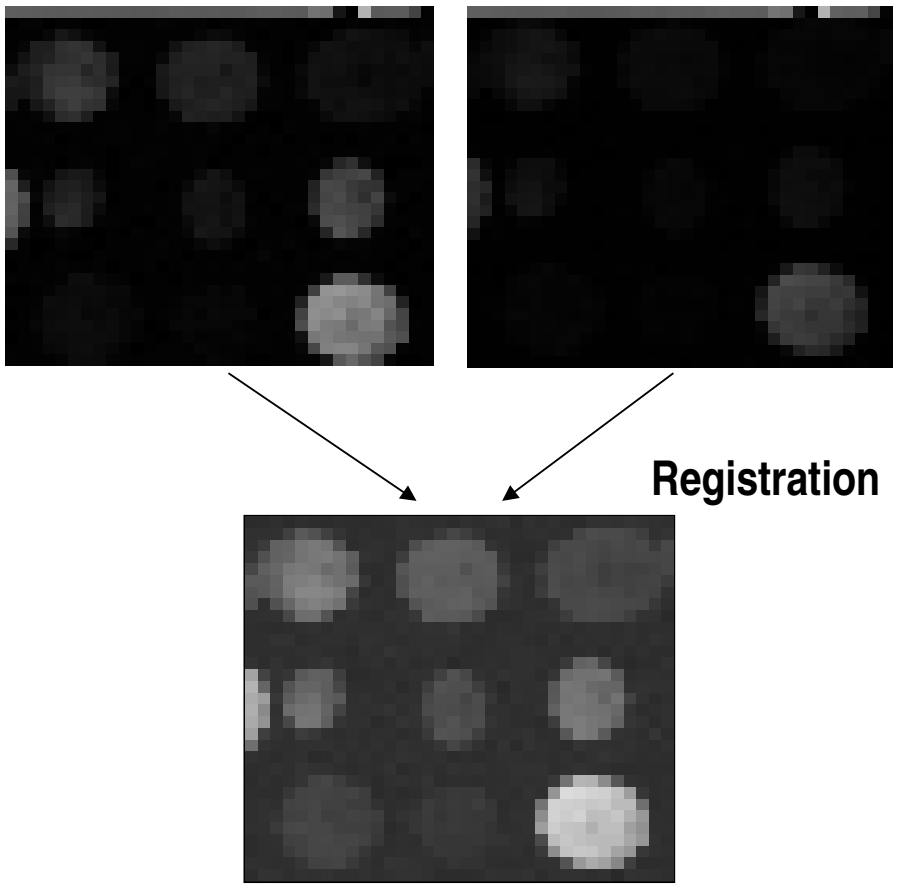


Terry Speed et al.



PASSO 1

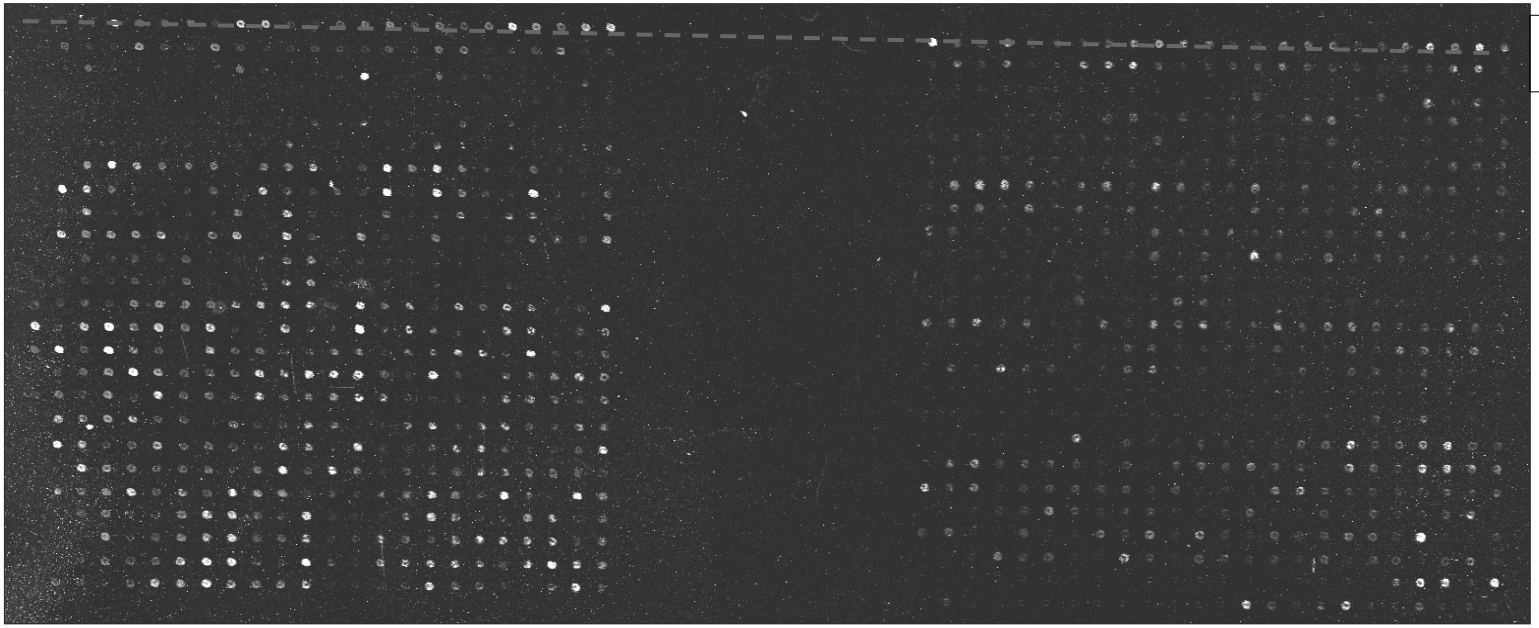
Identificare gli spot: assegnare una coordinata ad ognuno degli spot



Per gli array a due condizioni:
occorre registrare
l'informazione dei due array

Difficoltà:

- l'array può essere ruotato
- l'array può avere una distorsione prospettica
- per gli array a due dimensioni potrebbe esserci un disallineamento tra I due canali



PASSO 2: Preprocessing (normalizzazione) e Segmentazione

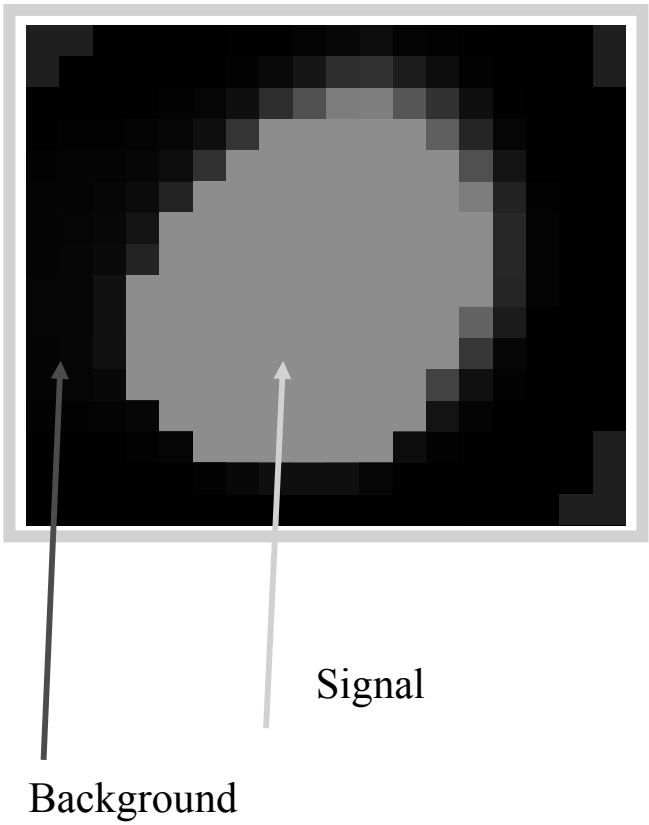
- Normalizzazione: per eliminare la variabilità derivante dalle condizioni sperimentali:
  - Scanning (laser and detector, chemistry of the flourescent label))
  - Hybridization (temperature, time, mixing, etc.)
  - Probe labeling
  - RNA extraction
  - Biological variability

- Approccio tipico: portare tutte le immagini alla stessa intensità di colore media

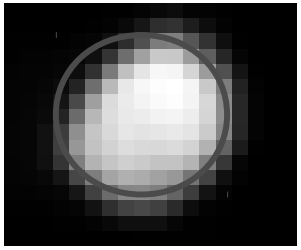
Problemi di Image processing

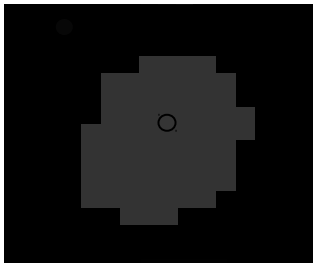
SEGMENTAZIONE

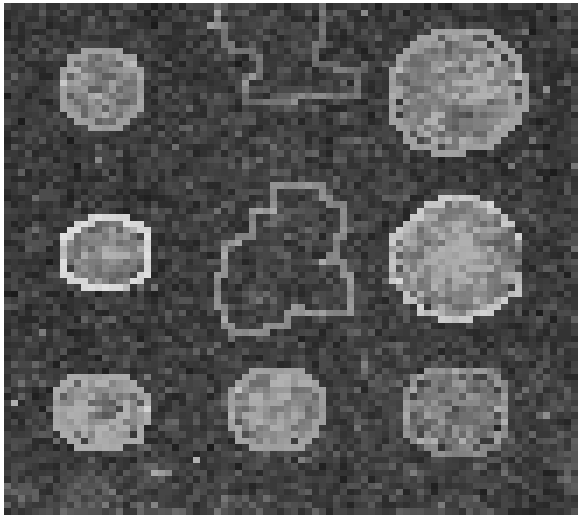
Classificazione di pixel come segnale o background



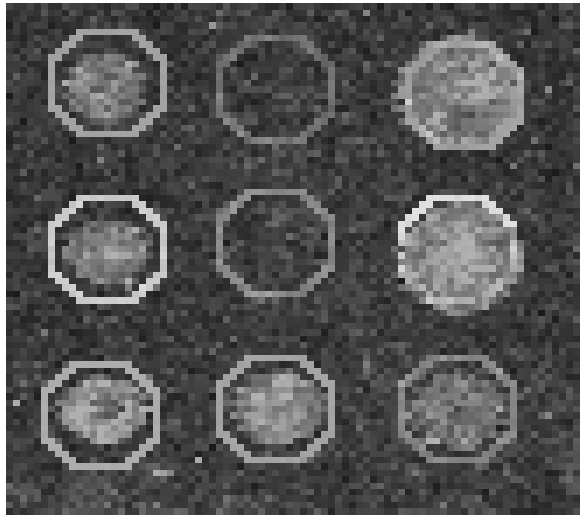
- Approcci:
  - Fixed circle
  - Adaptive Circle
  - Adaptive Shape (Region Growing)
  - ...







SRG



Fixed Circle

Problemi di Image processing

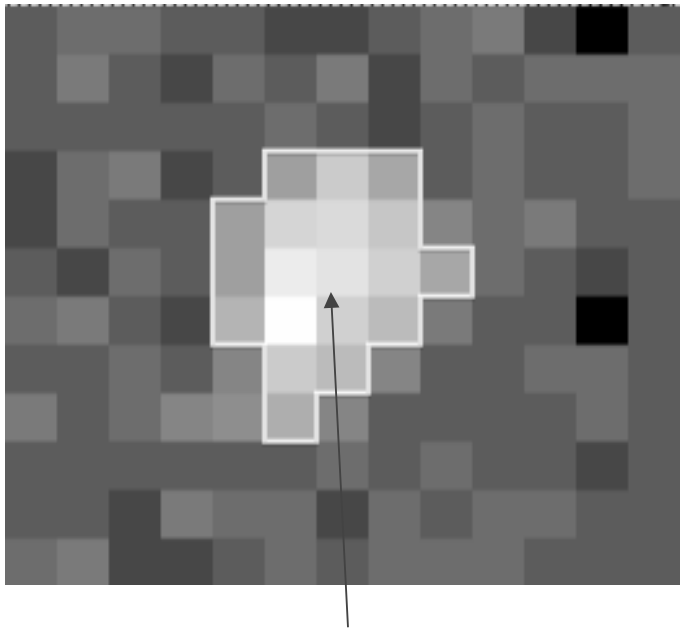
PASSO 3:

Quantificazione del segnale

a) stima del foreground: intensità media, intensità mediana

b) stima del background

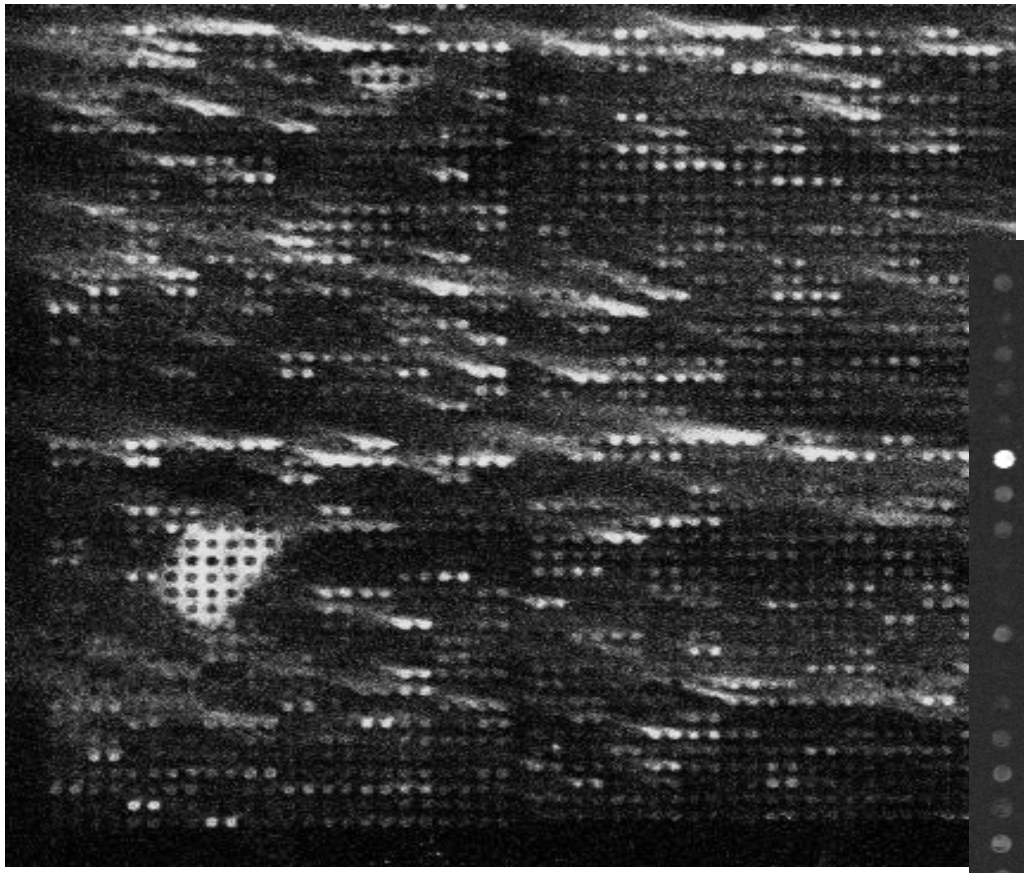
c) ... (shape, size etc)

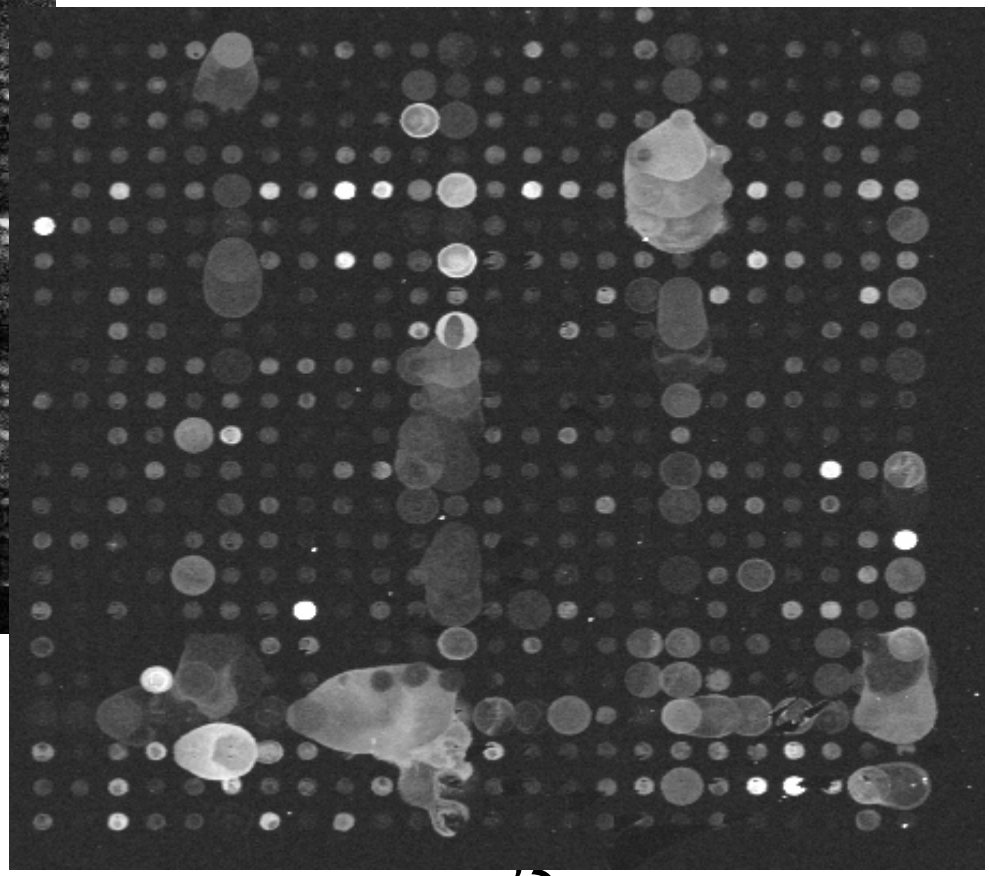


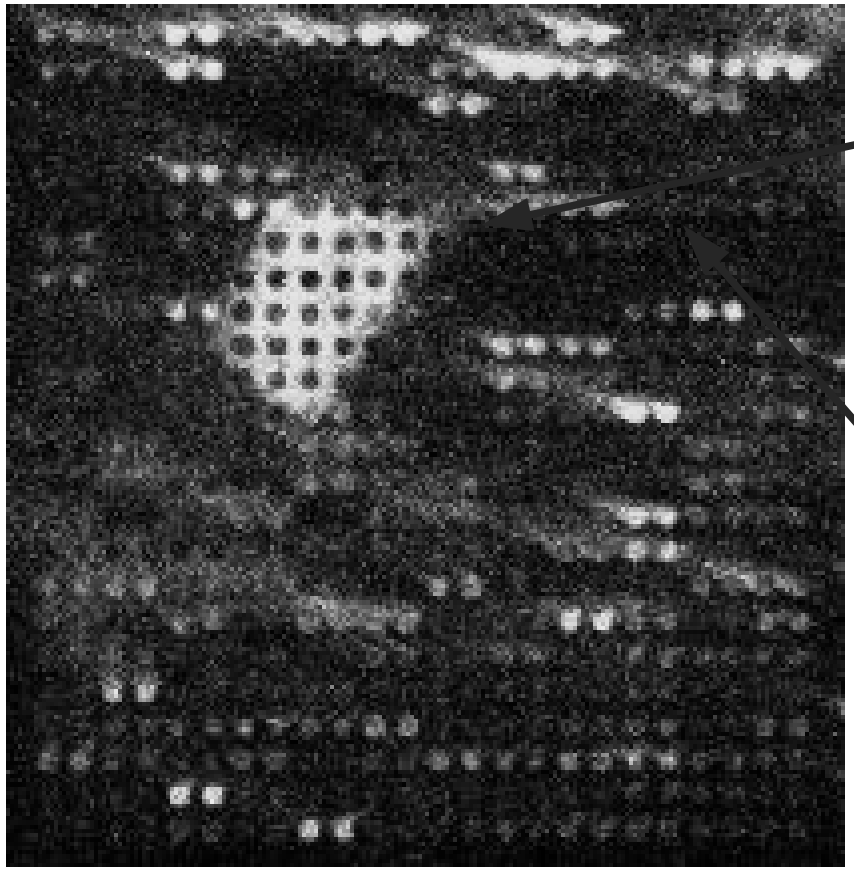
Media dei livelli di grigio

Terry Speed et al.

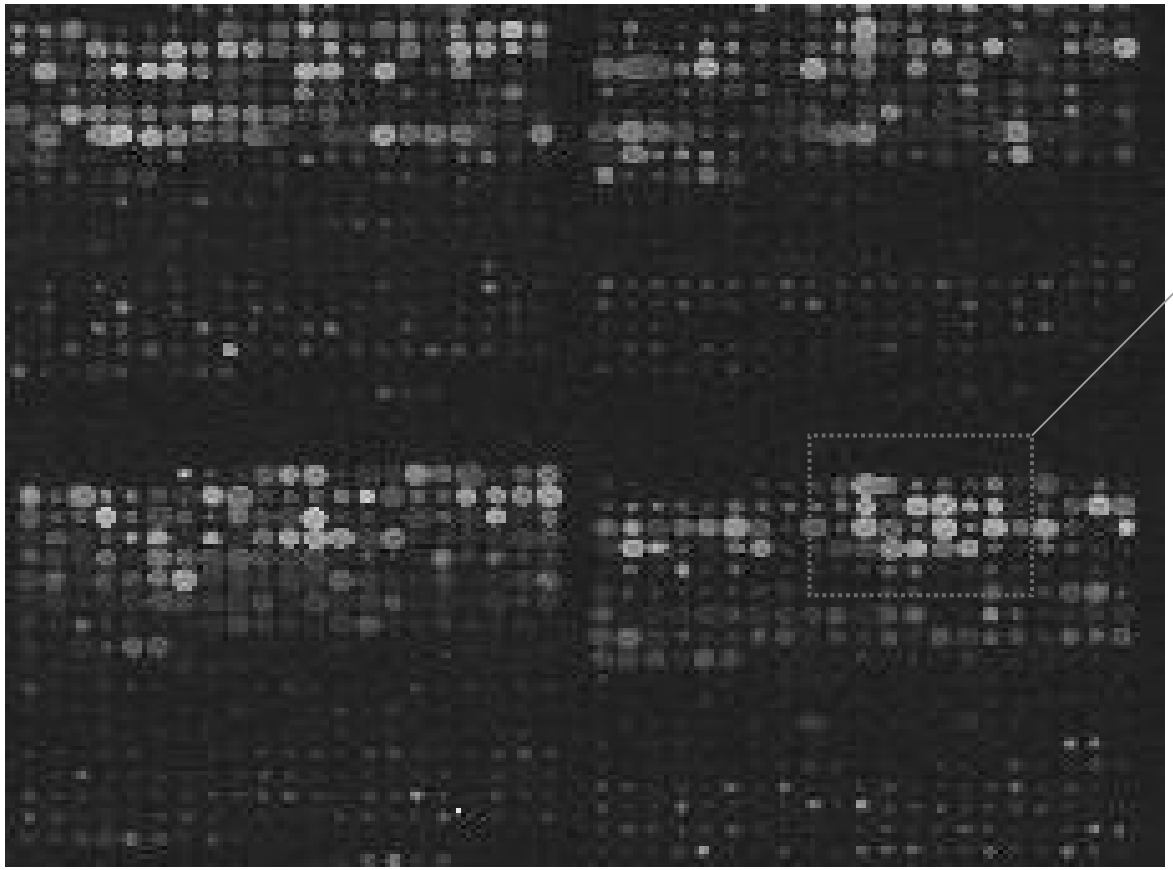
PASSO 4: stima della qualità dell'array e dei dati in generale

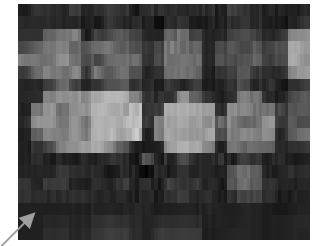


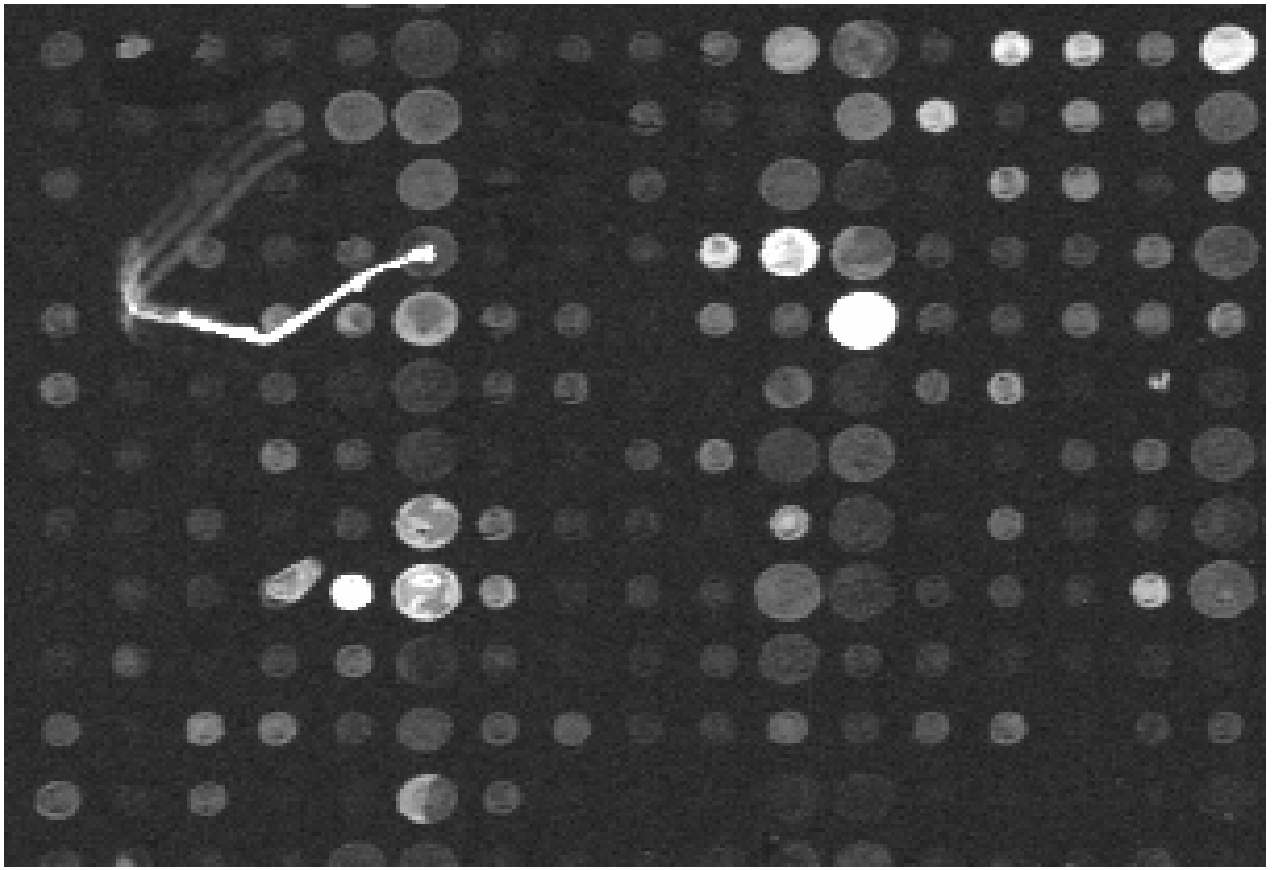



High Background
Weak Signals

Spot overlap:



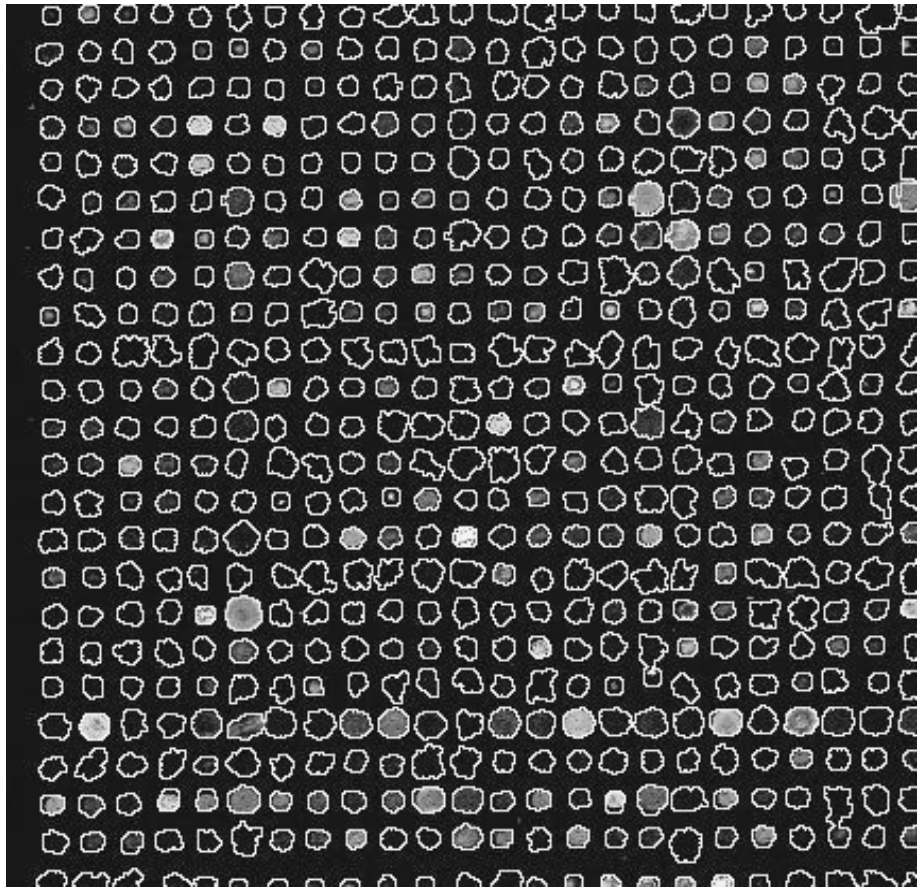




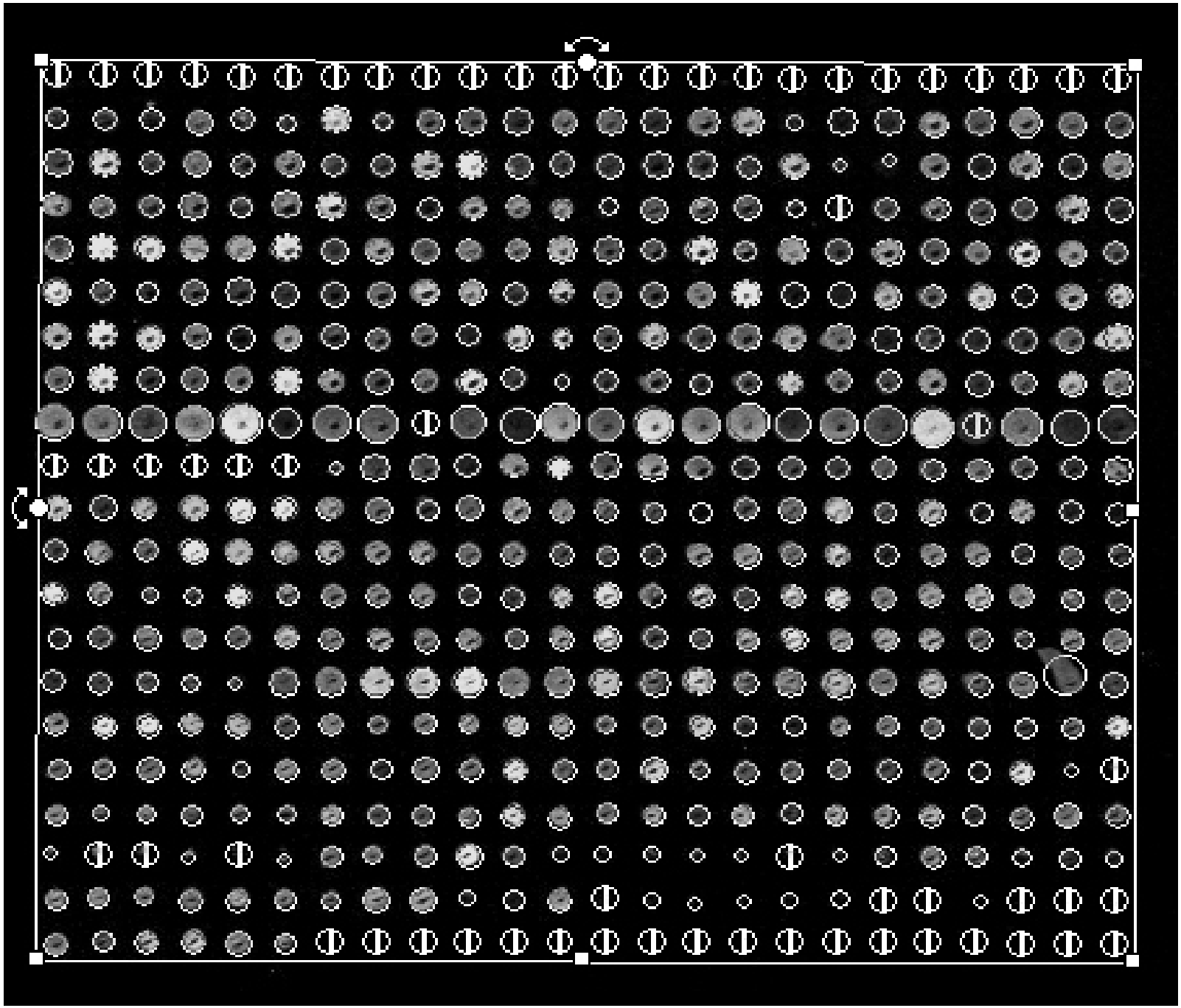
Riassumendo

1. Identificare gli spot
2. Preprocessing (normalizzazione) e Segmentazione
   - Classificazione di pixel come segnale o background
3. Quantificazione del segnale

   a) stima del foreground
   b) stima del background
   c) ... (shape, size etc)
5. Qualita’



Terry Speed et al.

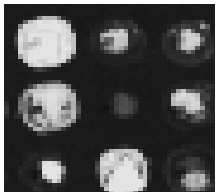


- La Pattern Recognition può essere utile anche in questa prima fase
  - Segmentazione degli spots: può essere affrontato con tecniche di clustering (vedi la lezione sull'analisi di immagini biomedicali)
  - Qualità: si può automatizzare il sistema di valutazione della qualità degli spot
    - Vediamo un esempio

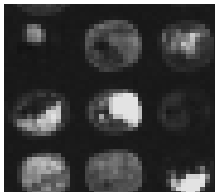
Stima della qualita’ degli spot con tecniche di Pattern Recognition

Problema:

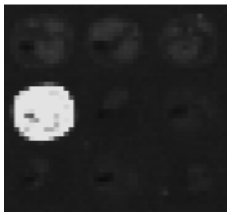
- ⇨ Rilevare gli spot di bassa qualita’



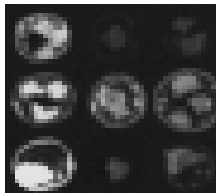
size



roundness



intensity



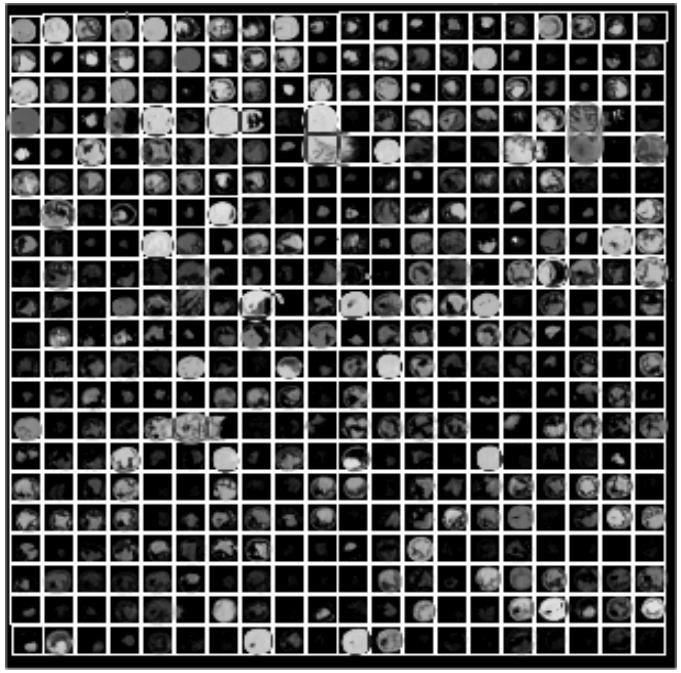
pixel distribution

Approccio tipico:

- ⇨ Annotazione manuale da parte degli esperti

L’approccio PR

Addestrare un modello utilizzando i giudizi degli esperti in un esperimento



spots (dati grezzi)

$\mathbf{x}_1$ y_1

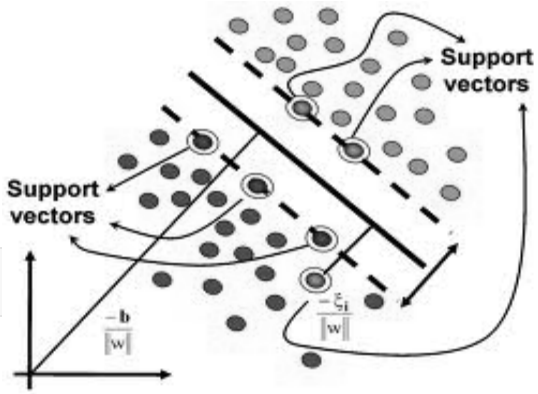
$\mathbf{x}_2$ y_2

… …

$\mathbf{x}_N$ y_N

features

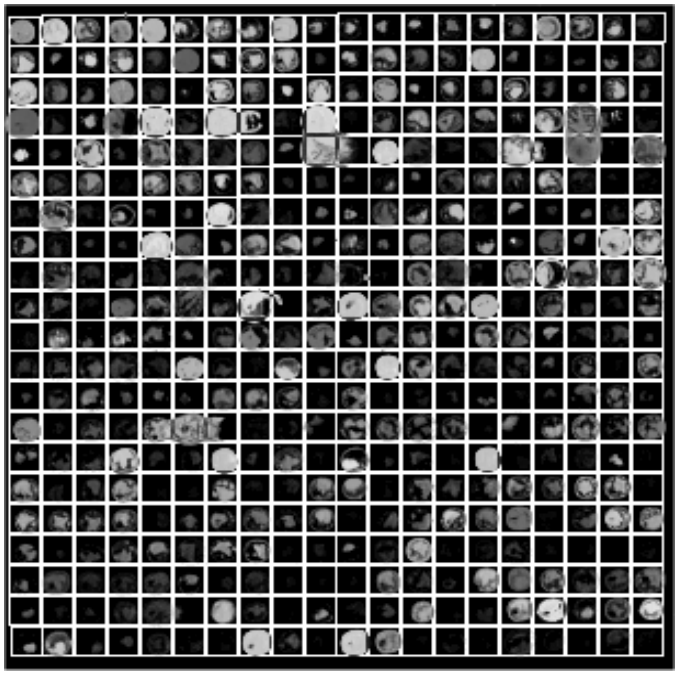
Etichette (buono/cattivo) date dagli esperti



Imparare come separare gli spot buoni da quelli cattivi

L’approccio PR

⇨ Testing: per qualsiasi esperimento



spots (dati grezzi)

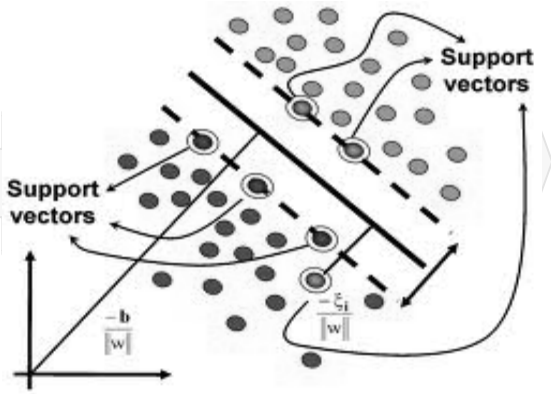
$\mathbf{x}_1$

$\mathbf{x}_2$

...

$\mathbf{x}_N$

features



Modello addestrato (senza l’intervento dell’esperto)

Per ogni spot: buono o non buono

La metodologia

- Features estratte: fittare una Gaussiana sullo spot e calcolare
  - L’intensita’ dello spot
  - L’intensita’ del background
  - L’errore di allineamento
  - La rotondita’ dello spot
  - La dimensione dello spot
  - …
- Confronto tra diverse tecniche di classificazione
  - In particolare, utilizzo delle Support Vector Machines con Kernel rbf

Gli esperimenti

- Risultati sperimentali
  - Dataset di 155 spots (97 sono buoni)
  - Etichette date da tre esperti (etichetta finale presa a maggioranza)
  - Accuratezza calcolata con la cross validation

Method	Accuracy
B-Course (subjective)	96.8%
Pair-wise NB (subjective)	95.5%
NB (subjective)	95.5%
NB (uniform)	94.8%
Decision Tree	91.6%
Neural Networks	90.3%
The proposed approach	97.4%

Problematiche di pattern recognition nell'analisi di dati di Espressione genica

Nota preliminare: condizioni multiple

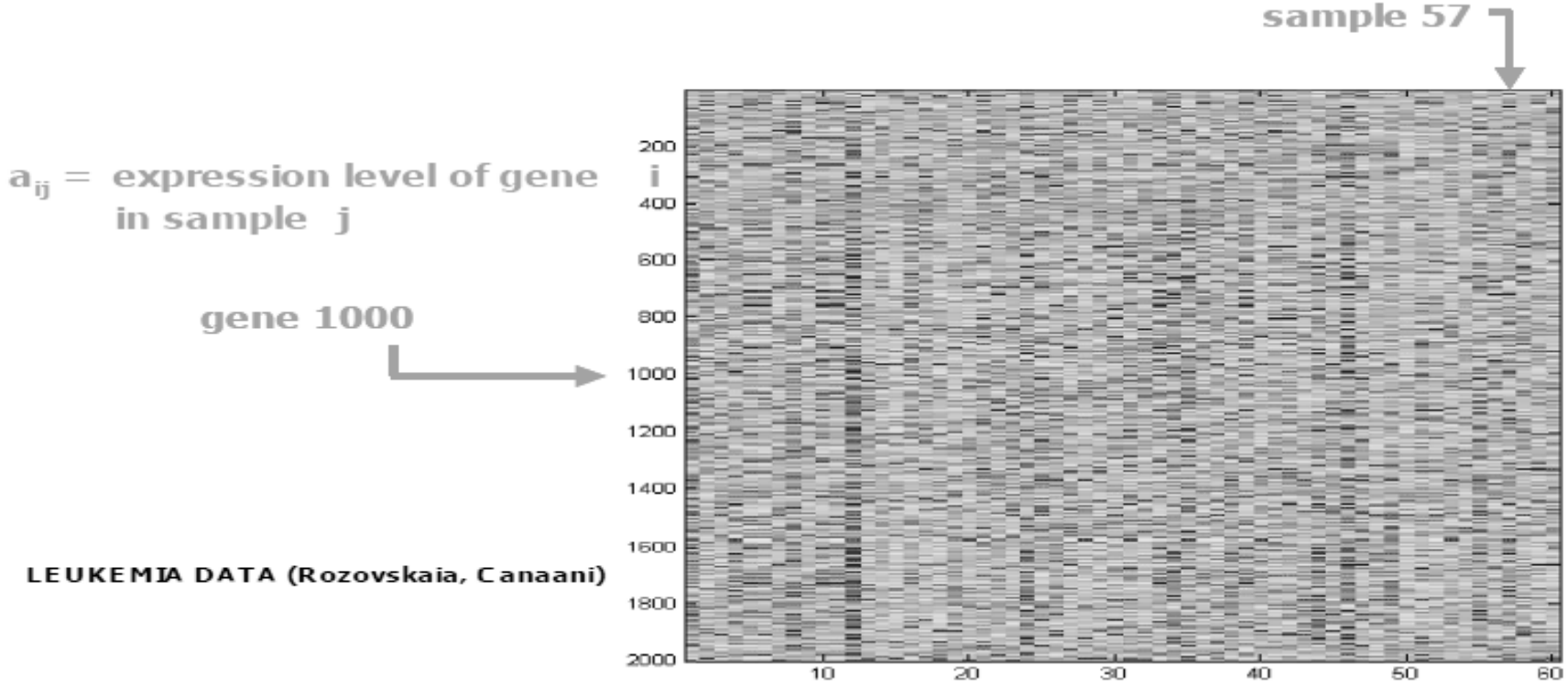
- Si può misurare il livello di espressione di un grande numero di geni in una serie di condizioni sperimentali differenti (campioni)

- I campioni possono corrispondere a:
  - Differenti istanti di tempo
  - Differenti condizioni ambientali
  - Differenti organi
  - Tessuti sani o malati
  - Diversi individui

Condizioni multiple

- L'espressione dei geni viene sistemata in una matrice di dati, dove:
  - Ogni gene corrisponde ad una riga
  - Ogni condizione corrisponde ad una colonna

- Ogni elemento della matrice rappresenta il livello di espressione di un gene in una specifica condizione
  - E' rappresentato da un numero reale che tipicamente e' il logaritmo dell'abbondanza relativa di mRNA del gene sotto la specifica condizione



Four conditions

Problematiche di PR

Trovare i geni che cambiano espressione tra campioni e controlli (“analisi statistica”)

Classificare i campioni sulla base del profilo di espressione dei geni (“classificazione”)

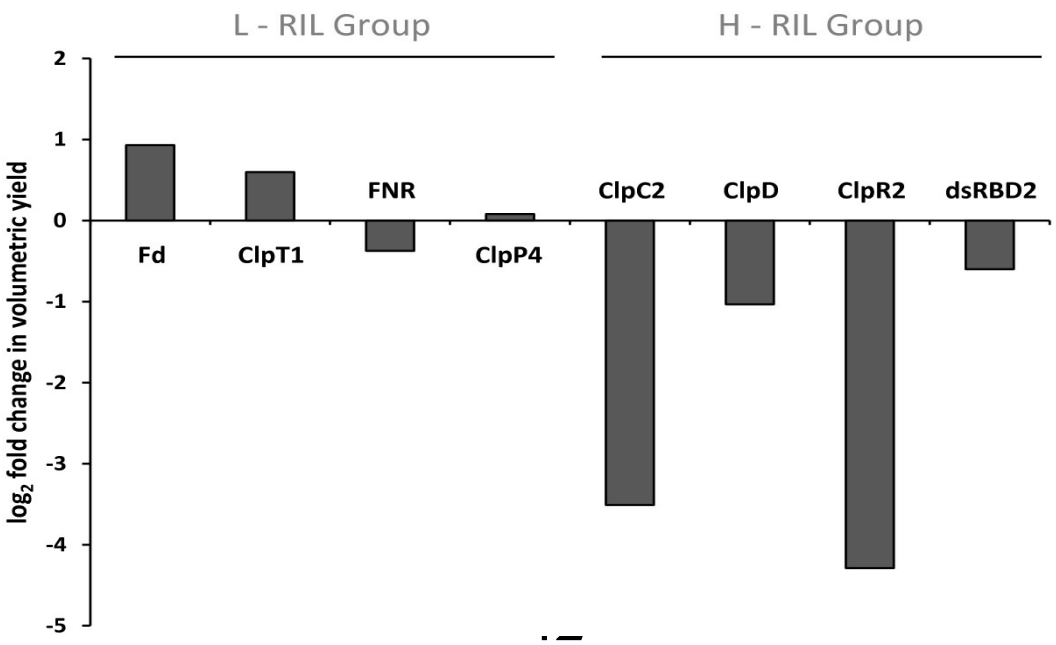
Clustering (di geni o di campioni): gruppi di geni o di campioni con comportamenti coerenti (“clustering”)

Trovare i geni che cambiano espressione tra campioni e controlli.

Approccio standard per calcolare l'aumento o la diminuzione dell'intensità di un gene in un campione rispetto al controllo:

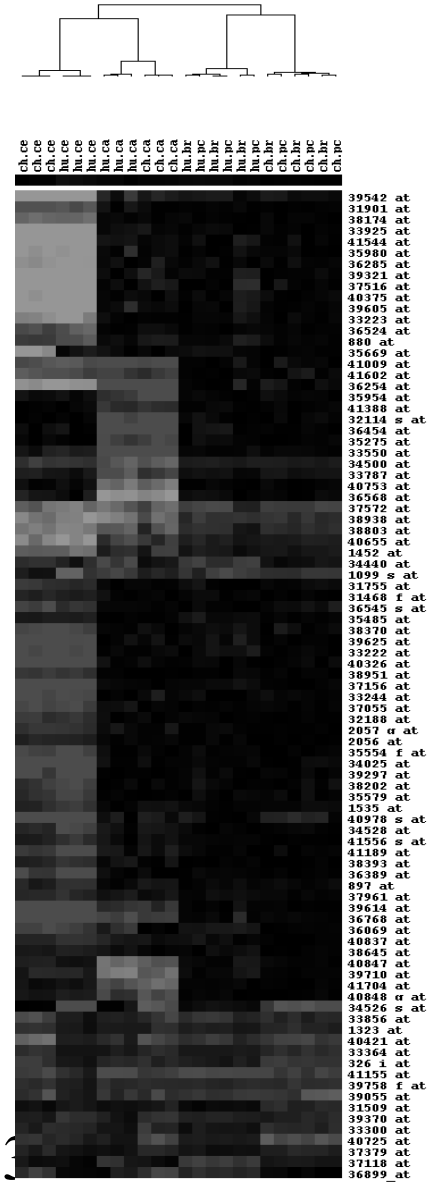
- Fold change
- t-statistics

NOTA: Occorre settare un cutoff per valori bassi (background +noise)



Classificazione di dati di espressione

- Goal: classificare diversi esperimenti sulla base dell'espressione genica
  - Distinguere tra sani e malati
- Problema difficile:
  - Rumore
  - Variabilità negli esperimenti
  - Variabilità biologica
  - Ridondanza tra geni
  - Curse of dimensionality! Pochi esperimenti, molti geni
    - Soluzione: gene selection

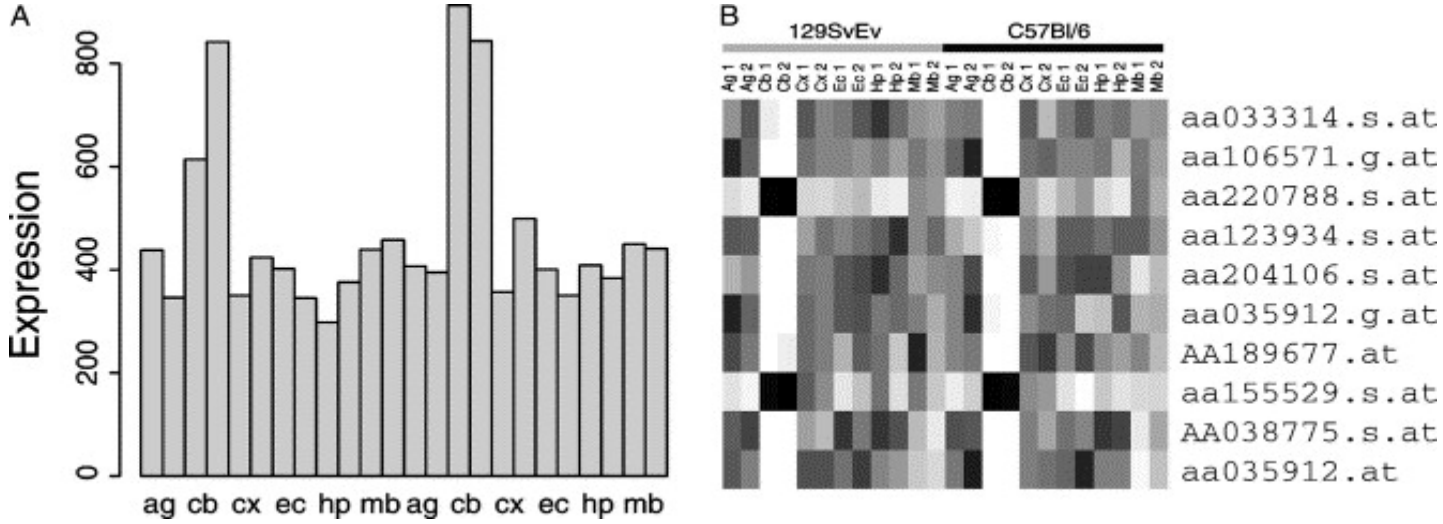


Gene selection

- “gene selection” è un processo mediante il quale si va a scegliere un gruppo ristretto di geni, ritenuti più significativi di altri in base al profilo di espressione (per esempio per discriminare tra condizioni sperimentali diverse).
- Due approcci
  - NON SUPERVISIONATI: non si tiene conto del problema
    - ⇨ VANTAGGI: semplici e veloci computazionalmente, indipendenti dall'algoritmo di classificazione;
    - ⇨ SVANTAGGI: ignorano l'interazione con il classificatore; feature considerate separatamente (problemi di peggior classificazione rispetto ad altre tecniche);

Gene selection

- Esempio: selezione basata sulla varianza o sull'entropia
  - I geni a varianza minore vengono scartati, rimangono quelli che variano di più
  - (IDEA: se un gene non cambia valore nell'insieme degli esperimenti non è rilevante)

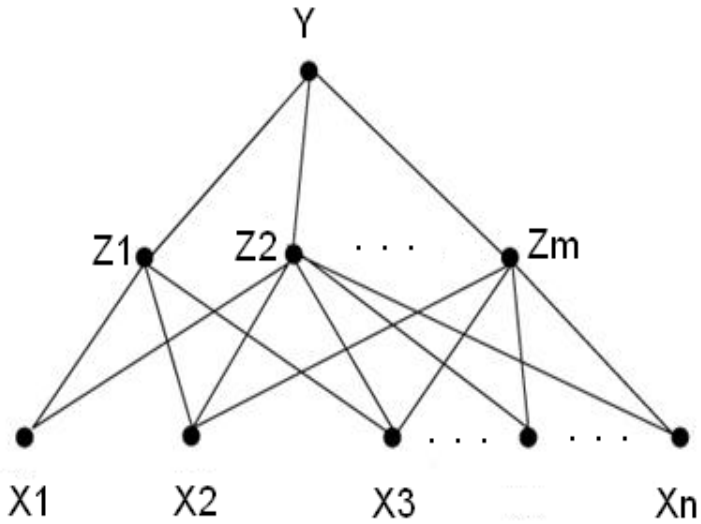


Gene selection

- SUPERVISIONATI: si tiene conto del problema (si usano le etichette)
  - l'utilità delle features è determinata dall'accuratezza stimata dall'algoritmo di learning;
  - VANTAGGI: interazione fra feature e modello;
  - SVANTAGGI: alto rischio di overfitting, alto costo computazionale;
  - ESEMPIO 1: Sequential Forward Feature Selection; si parte da un insieme vuoto di feature e progressivamente si aumenta il numero di feature da considerare, se massimizzano la prob. di corretta classificazione.

Gene selection

Esempio 2: Support Vector Machine – Recursive Feature Elimination



feature inputs or genes

$$\text{Traning set } \{x_k, y_k\}_{k=1}^{N}$$

$$\text{Input } x_k = (x_{k,1}, x_{k,2}, \text{K } x_{k,n})$$

$$\text{Ranking, } r_i = w_i = \sum_{k=1}^{N} \alpha_k y_k x_{k,i}$$

Clustering di dati di espressione

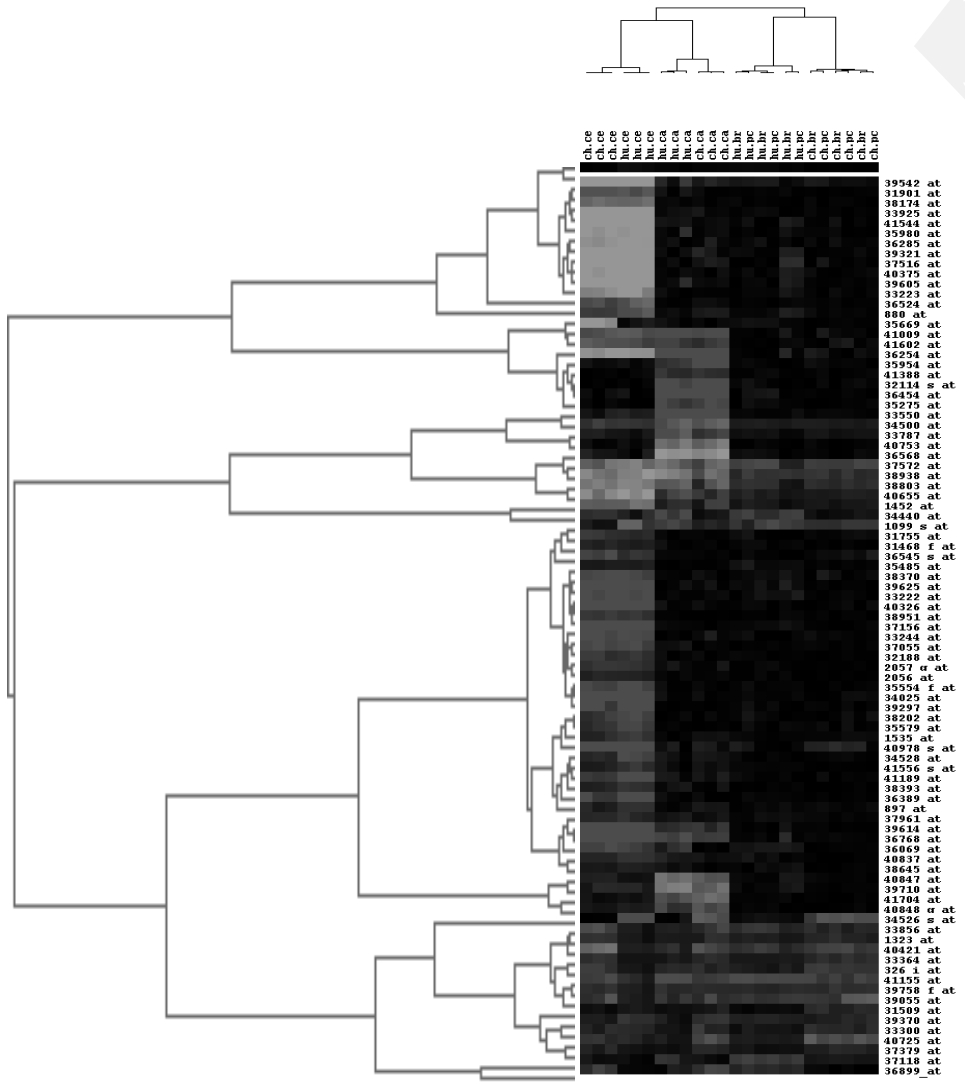
- Estrazione di informazioni utilizzando metodologie di clustering
- L’idea e’ quella di scovare similarita’ tra diversi livelli di espressione, in modo da determinare gruppi di geni o condizioni con comportamenti simili
  - Clustering di geni o clustering di condizioni
  - Applicazione delle diverse tecniche (da vedere nella parte sul clustering)

clustering di condizioni

Trovare esperimenti con geni espressi in modo simile: utile per identificare nuove classi (o sottoclassi) di malattie

clustering di geni

Trovare geni con pattern di espressione simile nei diversi esperimenti – identificazione di geni co-regolati o gene networks



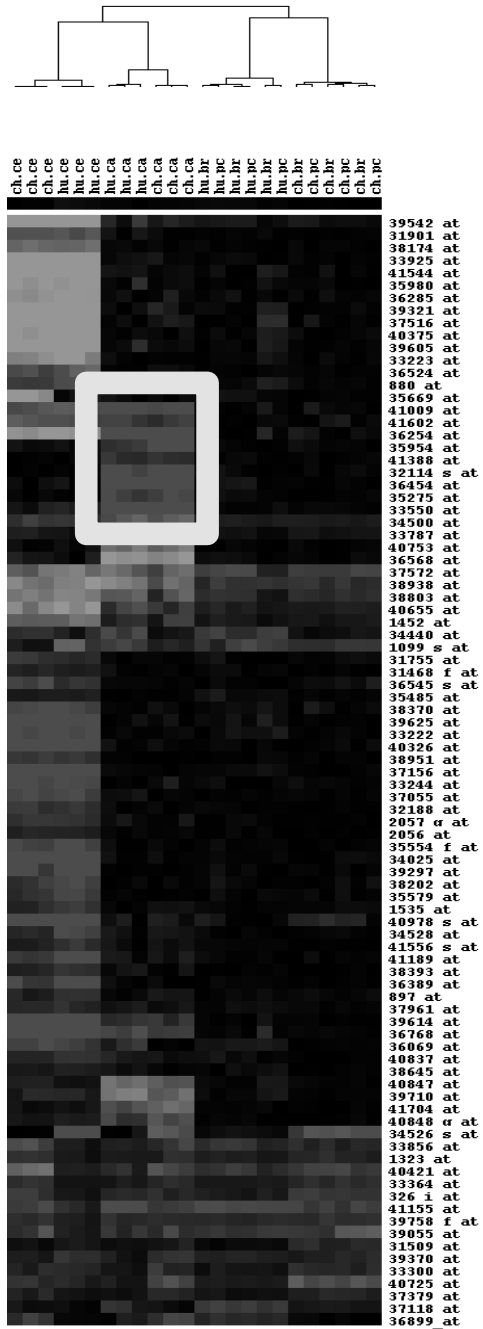
Commenti

- Clustering tra esperimenti: tipicamente poco utilizzato (di solito si hanno già tutte le informazioni necessarie)
- Tecniche utilizzate: approcci gerarchici, in grado di mettere in relazione i diversi geni

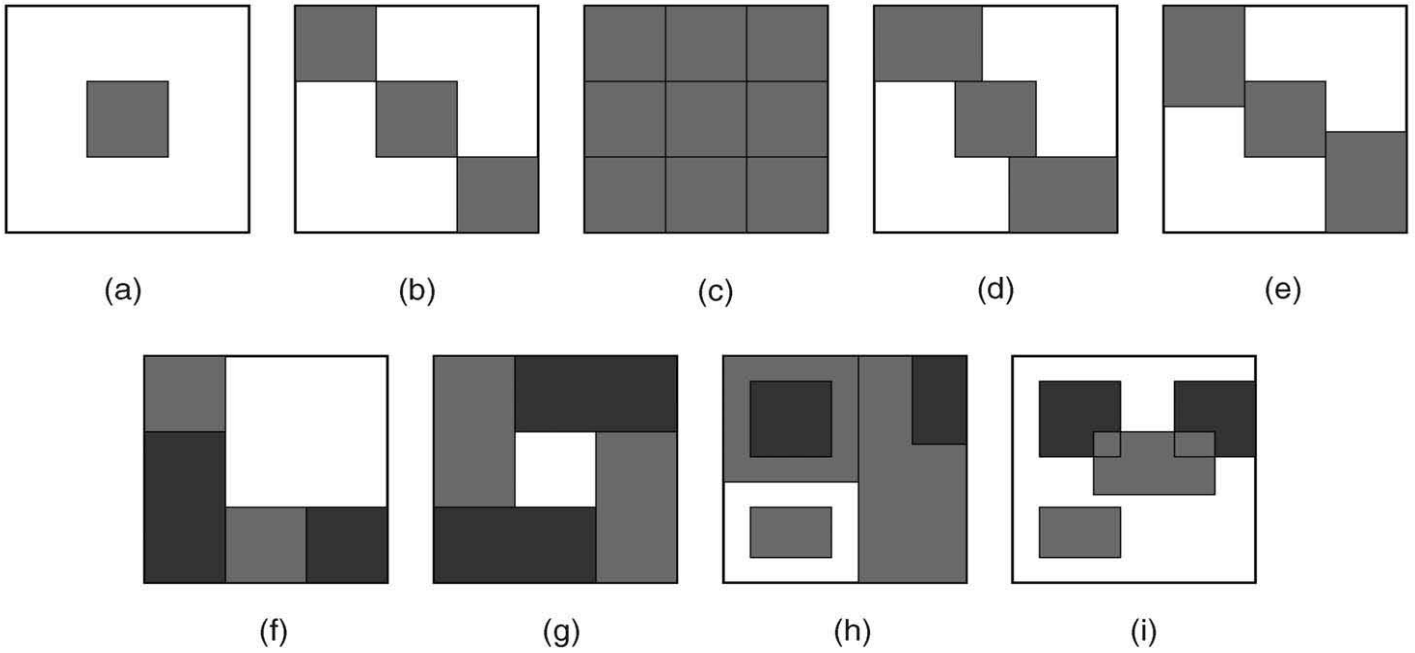
- PROBLEMA: i geni possono essere correlati solo in un sottoinsieme di esperimenti (ad esempio i geni “responsabili” di una certa malattia)

Biclustering

- Bicluster: sottoinsieme di geni che mostrano un comportamento “coerente” in un sottoinsieme di esperimenti
- Importante perchè ad un bicluster si potrebbe associare un processo biologico
  - Attivo solo in alcuni esperimenti (ad esempio solo nei malati)
  - Che coinvolge solo alcuni geni
- Problema complesso!



Possibili biclusters



(a) single bicluster
(b) exclusive row and column biclusters
(c) checkerboard structure
(d) exclusive rows biclusters
(e) exclusive columns biclusters

(f) nonoverlapping biclusters with tree structure
(g) nonoverlapping nonexclusive biclusters
(h) overlapping biclusters with hierarchical structure
(i) arbitrarily positioned overlapping biclusters

Un modello probabilistico per dati di espressione

- Idea: utilizzare un modello probabilistico utilizzato nel campo della linguistica: i topic models

- Vediamo:
  - Il punto di partenza per la linguistica: bag of words
  - Topic models per l'analisi di documenti
  - Il parallelismo documento / esperimento di espressione genica

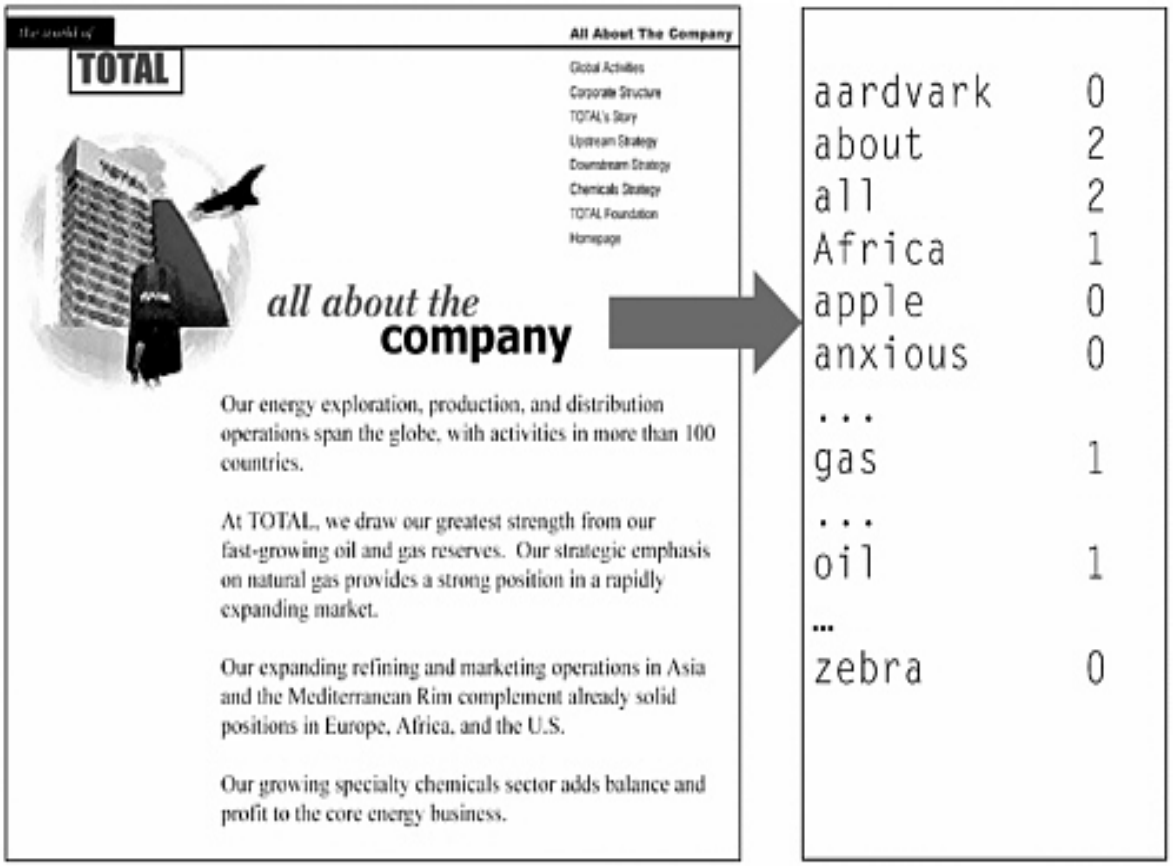
Bag of words

- Un documento viene visto come una collezione “non ordinata” di parole (una “bag of words”)
- Dato un dizionario, si può riassumere un documento indicando il numero di volte che ogni parola compare in esso

Si è interessati al numero di occorrenze di ogni parola, piuttosto che alla posizione

Bag of words

- Ogni documento è caratterizzato da un “istogramma” di parole (un vettore lungo quanto il dizionario)



⇨ Problema: una parola può avere diversi significati a seconda del contesto

Sun?

Sun microsystems

Windows?

"Home"	"sports"	"space"	"computers"	"weather"
Kitchen	Team	Space	Drive	Rain
Door	Game	Sun	Windows	Snow
Garden	Play	Research	Card	Sun
Windows	Year	Center	DOS	Season
Bedroom	Games	Earth	SCSI	Weekend
Space	Season	NASA	Sun	Cloudy

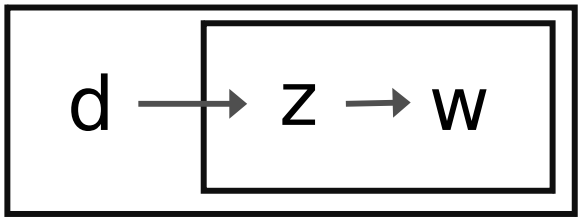
Soluzione: Topic Models

- L'idea è che le parole possono essere disambiguate guardando al contesto
- I topic models introducono un livello intermedio, basato sul concetto di “topic” (argomento)
  - Rappresenta il concetto di “Di cosa stiamo parlando?”
  - I topics sono estratti in modo automatico guardando alla co-occorrenza delle parole nei vari documenti

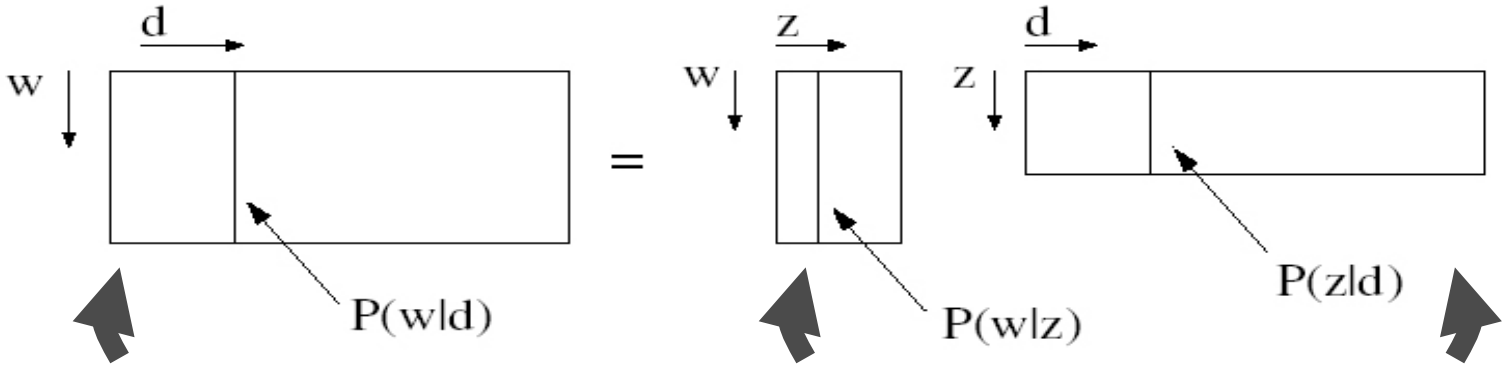
In altre parole:

- Ogni documento può essere caratterizzato dalla presenza di diversi topic
  - Esempio: un articolo della gazzetta dello sport parla al 60% del topic “calcio” e al 40% del topic “economia”
- Ogni topic induce un particolare set di parole
  - Esempio: se si parla di “calcio” è molto probabile trovare le parole “stadio”, “partita”, “allenamento”, ...

- Un esempio di topic model è la pLSA (probabilistic Latent Semantic Analysis)
- Punto di partenza:
  - Una collezione di documenti descritti da una matrice n(w,d)
  - n(w1,d1) indica il numero di occorrenze della parola w1 nel documento d1
- La pLSA modella e descrive la probabilità di trovare una data parola in un documento
- Questa probabilità è mediata dai topics



$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j)$$



Observed word distributions

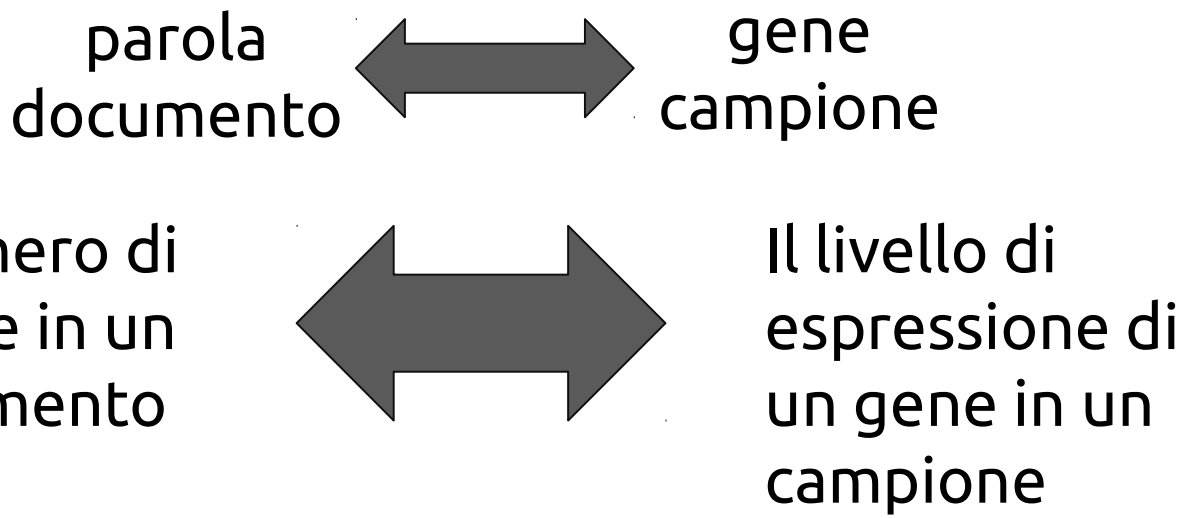
word distributions per topic

Topic distributions per document

- Training della pLSA:
  - Stimare le probabilità p(w|z) e p(z|d)
- Cosa possiamo fare una volta addestrata la pLSA
  - Dato un documento, possiamo capire “di quali argomenti si parla” (usando la p(z|d))
  - Dato un argomento, possiamo capire “quali sono le parole più legate a quell'argomento” (usando la p(w|z))

PLSA e Espressione genica

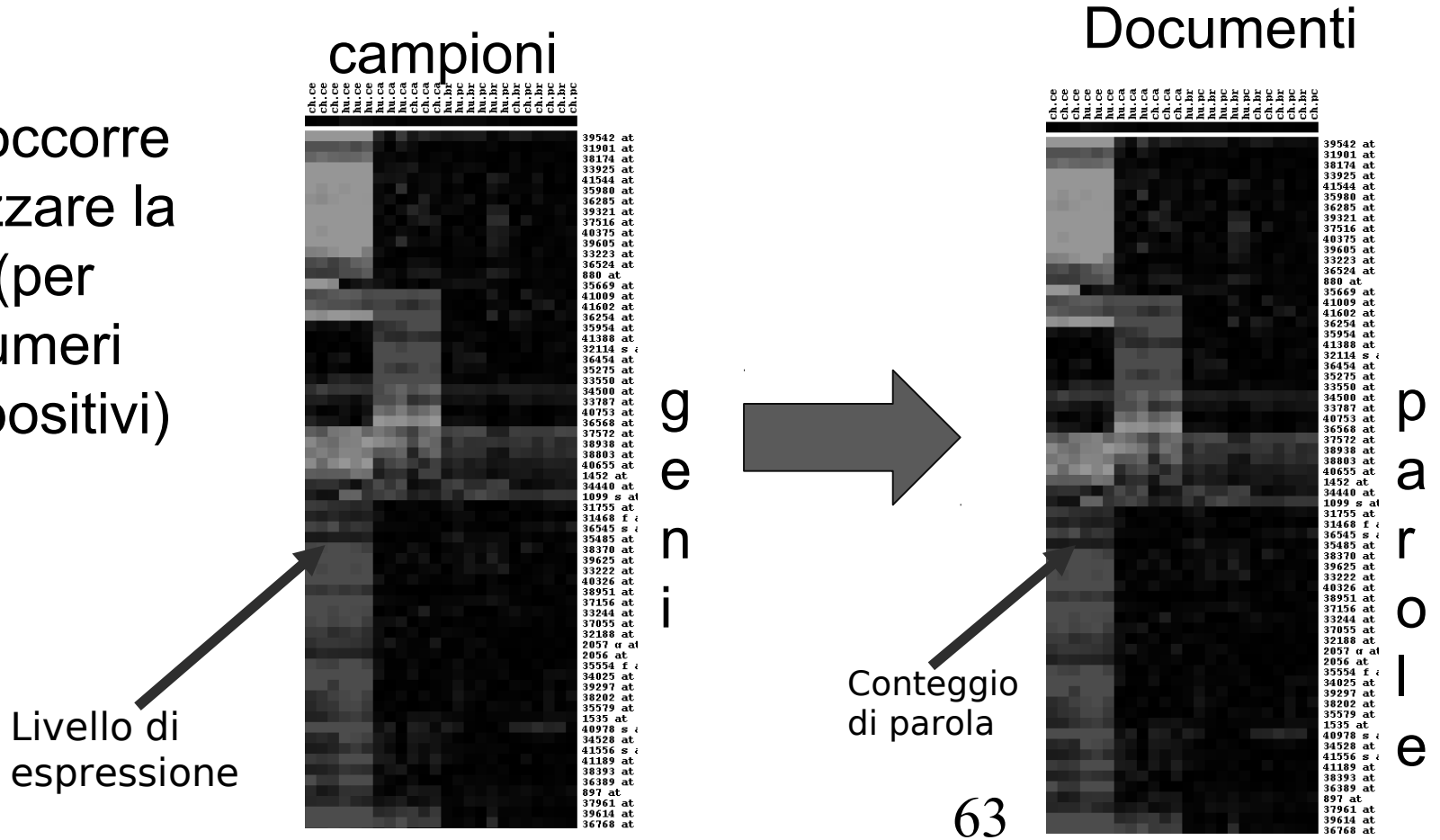
- Possiamo settare un'analogia tra l'analisi di documenti e l'analisi di dati di espressione
  - Un documento è caratterizzato dalla diversa presenza delle parole
  - Un esperimento è caratterizzato dal diverso livello di espressione dei geni



PLSA e Espressione genica

- Possiamo quindi applicare la pLSA ai dati di espressione considerando la matrice di espressione come la matrice dei counts (la n(w,d))

NOTA: occorre normalizzare la matrice (per avere numeri interi e positivi)



Plsa e Espressione genica

- Utile per classificazione: possiamo caratterizzare ogni esperimento con la sua distribuzione p(z|d) (“di che argomenti si parla”)
  - Dimostrato in altri contesti che questa rappresentazione è molto descrittiva e discriminante

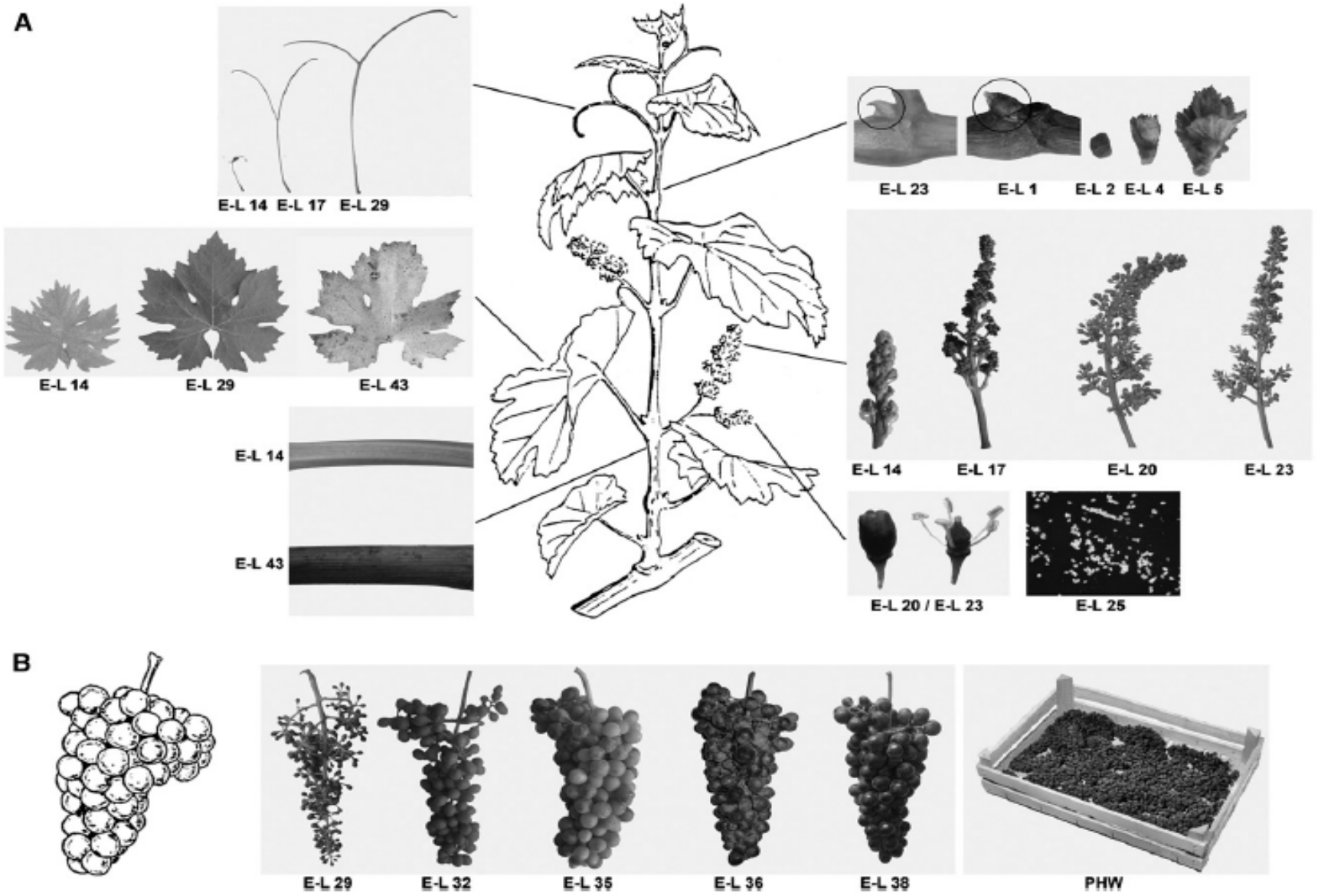
- Esempio: classificazione di immagini (documento → immagine, parola → feature visuale)

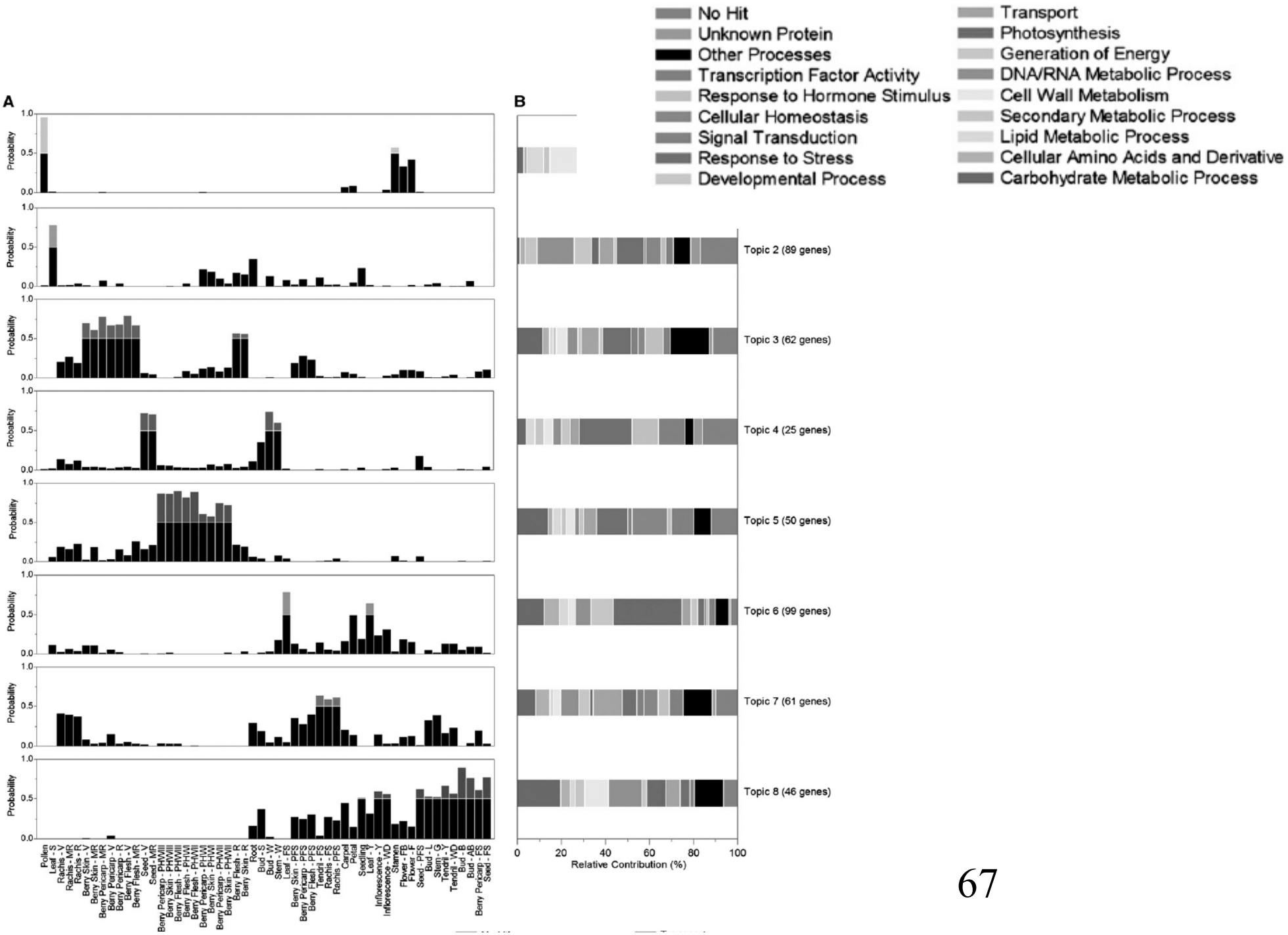
PLSA e Espressione genica

Feature importante: Interpretabilità

- Possiamo associare ad ogni topic un “processo biologico”
  - Attivo in determinati campioni (dove “si parla” di quel processo)
  - Che coinvolge particolari geni (i geni coinvolti in quel processo biologico)
- **P(z|d)**: può essere usata per capire quali sono (e in che misura) i processi attivi nei differenti campioni
- **P(w|z)**: può rappresentare l'impatto dei diversi geni nel particolare processo biologico

Esempio: ATLAS della vite




A
B
Probability
No Hit
Unknown Protein
Other Processes
Transcription Factor Activity
Response to Hormone Stimulus
Cellular Homeostasis
Signal Transduction
Response to Stress
Developmental Process
Transport
Photosynthesis
Generation of Energy
DNA/RNA Metabolic Process
Cell Wall Metabolism
Secondary Metabolic Process
Lipid Metabolic Process
Cellular Amino Acids and Derivative
Carbohydrate Metabolic Process
Topic 2 (89 genes)
Topic 3 (62 genes)
Topic 4 (25 genes)
Topic 5 (50 genes)
Topic 6 (99 genes)
Topic 7 (61 genes)
Topic 8 (46 genes)
Relative Contribution (%)
Pollen
Leaf - S
Rachis - V
Rachis - MR
Rachis - R
Berry Skin - V
Berry Skin - MR
Berry Pericarp - MR
Berry Pericarp - V
Berry Pericarp - R
Berry Flesh - V
Berry Flesh - MR
Seed - V
Seed - MR
Berry Pericarp - PHWII
Berry Skin - PHWII
Berry Flesh - PHWII
Berry Flesh - PHWI
Berry Pericarp - PHWI
Berry Skin - PHWI
Berry Pericarp - PHWII
Berry Skin - PHWII
Berry Flesh - R
Berry Skin - R
Root
Bud - S
Bud - W
Stem - W
Leaf - FS
Berry Skin - PFS
Berry Pericarp - PFS
Berry Flesh - PFS
Tendril - FS
Rachis - PFS
Rachis - FS
Carpel
Petal
Seedling
Leaf - Y
Inflorescence - Y
Inflorescence - WD
Stamen
Flower - FB
Flower - F
Seed - FFS
Bud - L
Stem - G
Tendril - Y
Tendril - WD
Bud - AB
Bud - B
Berry Pericarp - FS
Seed - FS

Un altro esempio: la resistenza ai patogeni