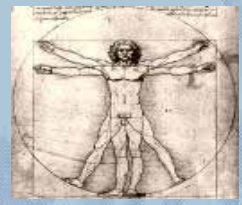
Biclustering of Espression Data Using Simulated Annealing

Kenneth Bryan, Pàdraig Cunningham, Nadia Bolshakova

Il genoma di molti organismi è stato sequenziato:



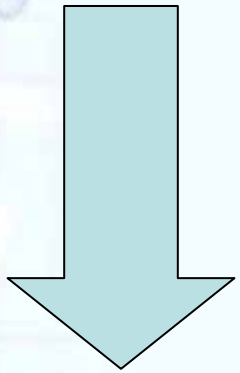


La maggior parte dei geni (> nell'uomo) ha funzione sconosciuta.

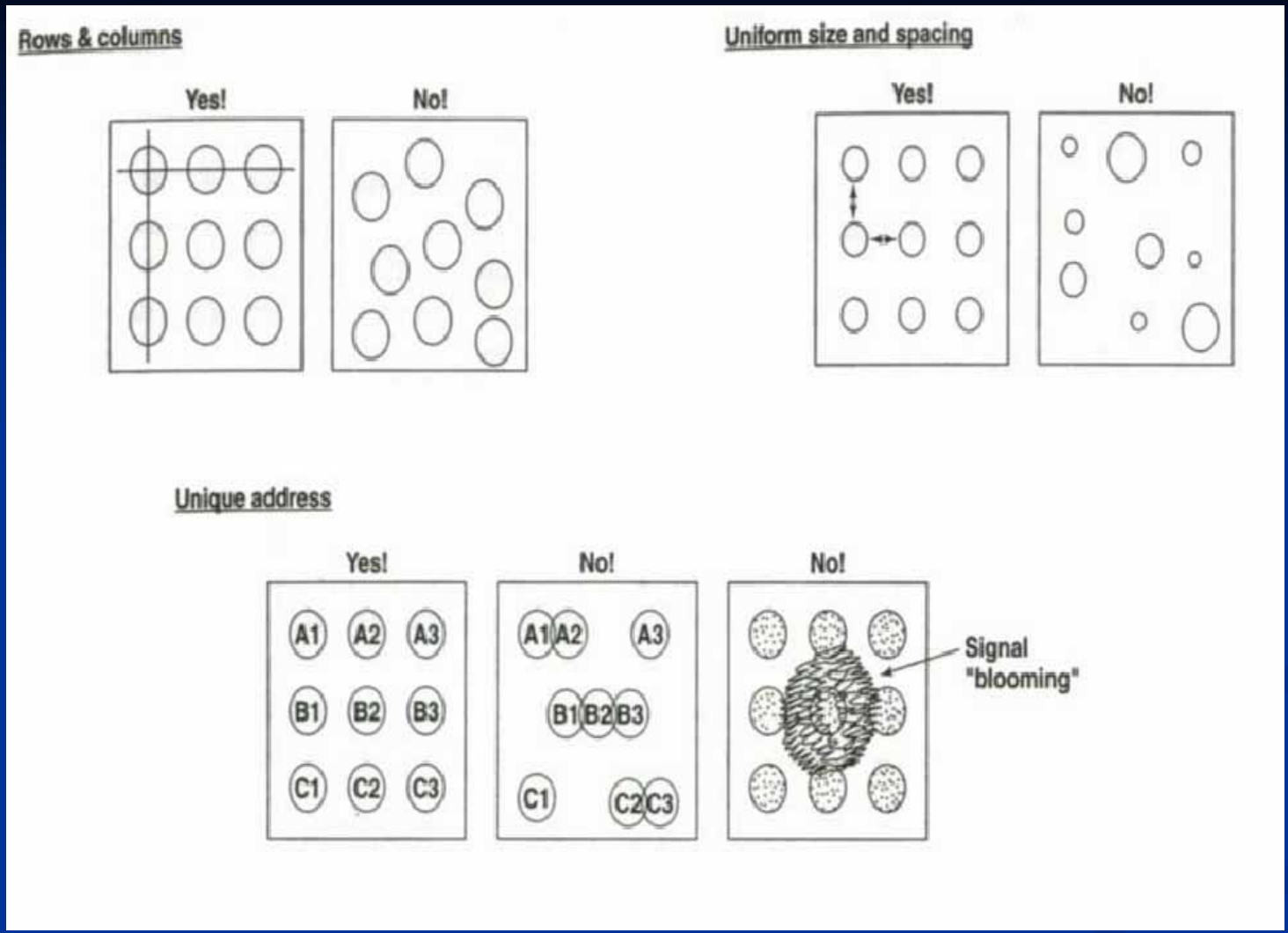
- **Conoscere la funzione di tutti i geni(studiando l'espressione genica)**
- **Raggruppare i geni in base alle funzioni**

Tecnologia microarray

MICROARRAY



matrice ordinata di elementi microscopici su un substrato planare, che consente il legame di specifici geni o di prodotti di geni.

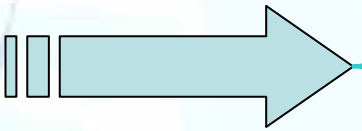


Rapida deposizione, individuazione e quantificazione degli spot

Substrari:

- **Vetro (favorisce il legame con la sonda)**
- **Materiali plastici**
- **Silicio**
- **Filtri di nylon**
- **Nitrocellulosa**

PLANARE

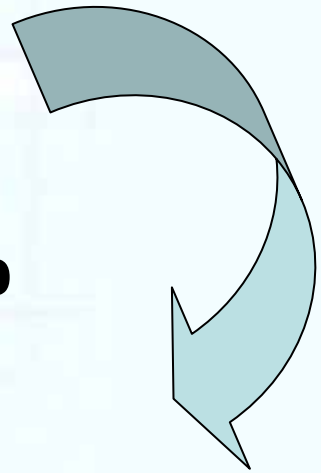


- Ottima automatizzazione della procedura di stampa tramite pin e ugelli ik-jet
- Accurato scattering grazie alla precisa individuazione della distanza tra gli elementi ottici dello scanner e la superficie

Vantaggio:

Monitorare l'attività di migliaia di geni contemporaneamente e in condizioni diverse:

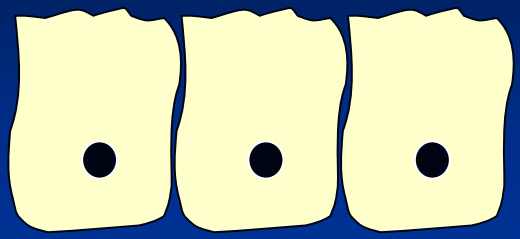
- **Tipo di tessuto**
- **Paziente**
- **Ambienti di sviluppo**



1. **Studio di tessuti**
2. **Scoprire le basi molecolari di una malattia(espressione genica, co-regolazione)**
3. **Analisi del meccanismo di azione dei farmaci**

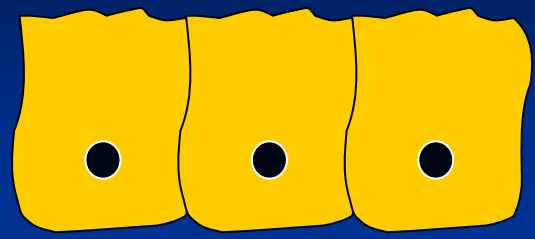
Gene A

Cellula normale



Basso livello di espressione

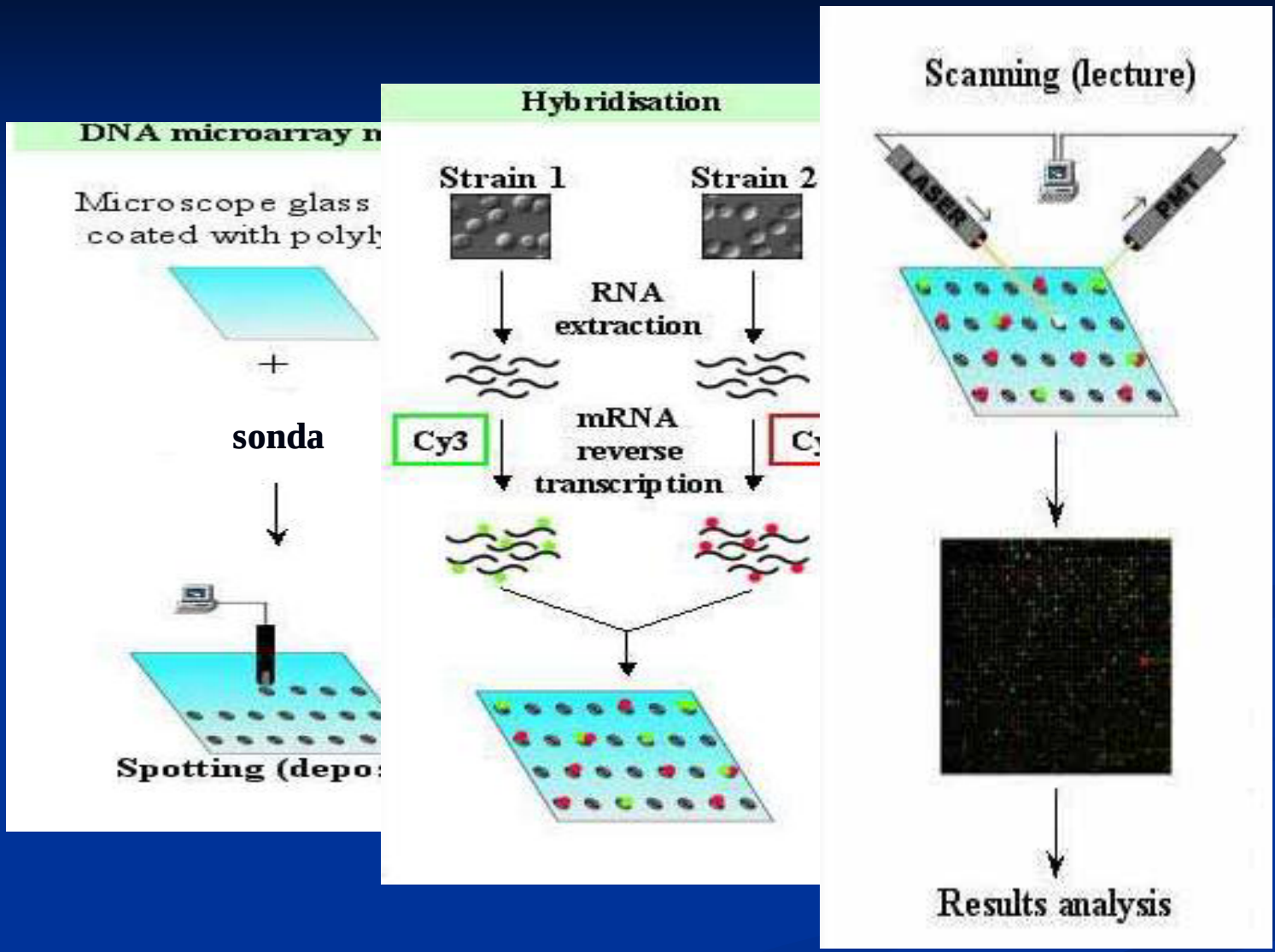
Cellula malata cancro



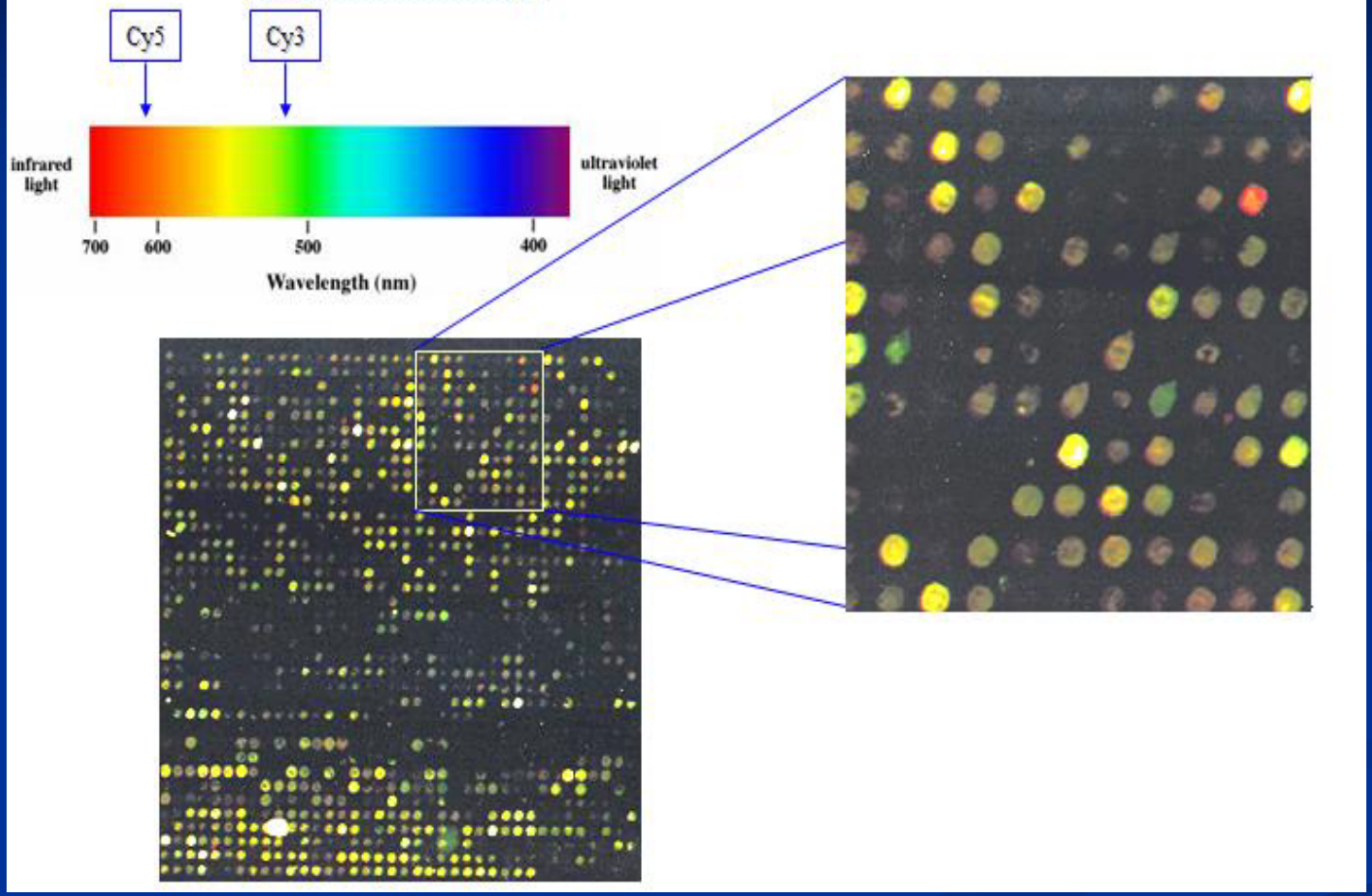
Alto livello di espressione

Il gene A potrebbe avere un ruolo importante nell'insorgenza del cancro!

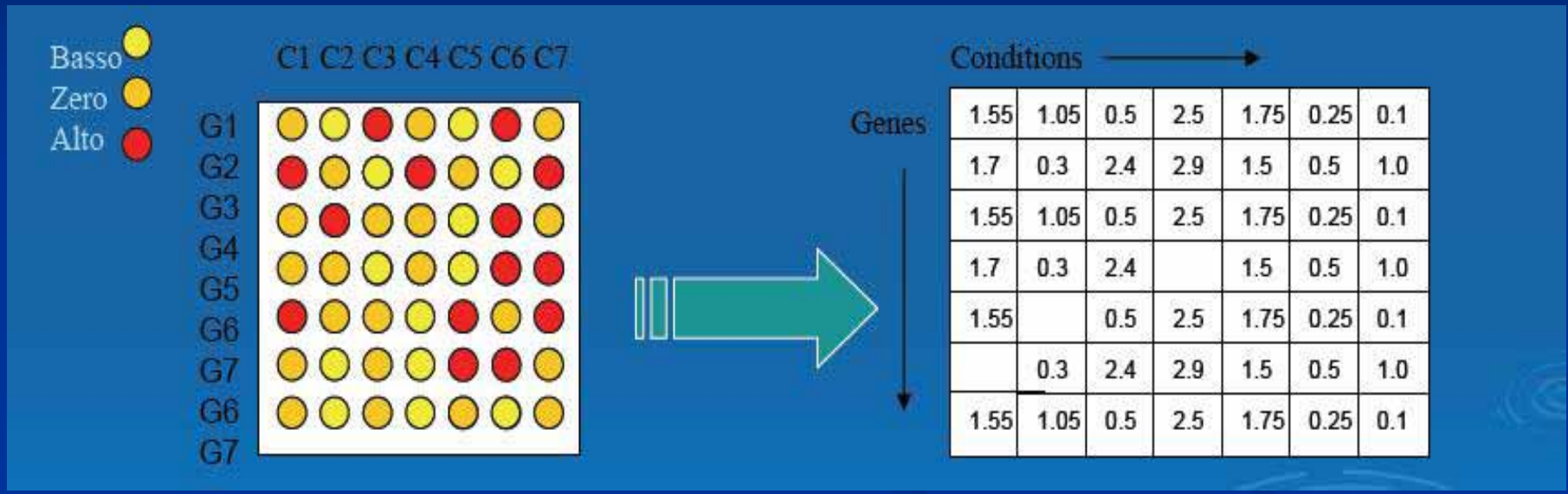
Preparazione standard : tecnica spotting



MicroArray con Fluorocromi



Gli esperimenti vengono eseguiti tramite una matrice simile a questa:



1.55	1.05	0.5	2.5	1.75	0.25	0.1
1.7	0.3	2.4	2.9	1.5	0.5	1.0
1.55	1.05	0.5	2.5	1.75	0.25	0.1
1.7	0.3	2.4		1.5	0.5	1.0
1.55		0.5	2.5	1.75	0.25	0.1
	0.3	2.4	2.9	1.5	0.5	1.0
1.55	1.05	0.5	2.5	1.75	0.25	0.1

I diversi colori indicano diversi livelli di espressione.

Per facilitare l'analisi computazionale la matrice viene convertita in una matrice numerica.

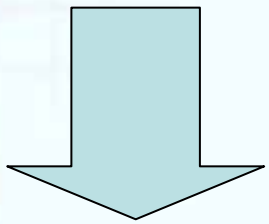
A (nxm)

n = numero di geni

m = numero condizioni (array)

a_{ij} **= livello di espressione del gene i nella condizione j**

Analisi molto difficile



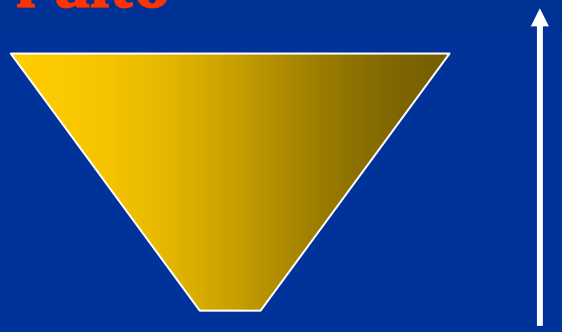
DATA MINING

(insieme di strumenti per estrapolare pattern)

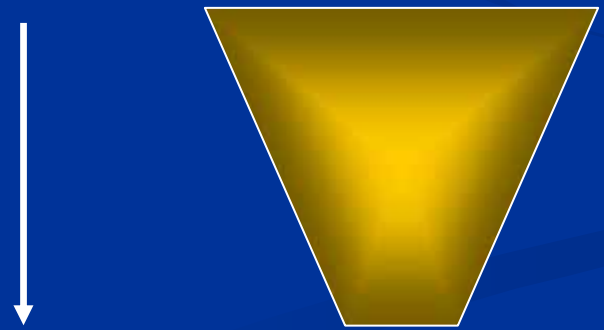
CLUSTERING

Tecnica di analisi di dati per selezionare e raggruppare elementi omogenei, basato sul concetto di distanza tra due elementi.

a) Basso verso l'alto



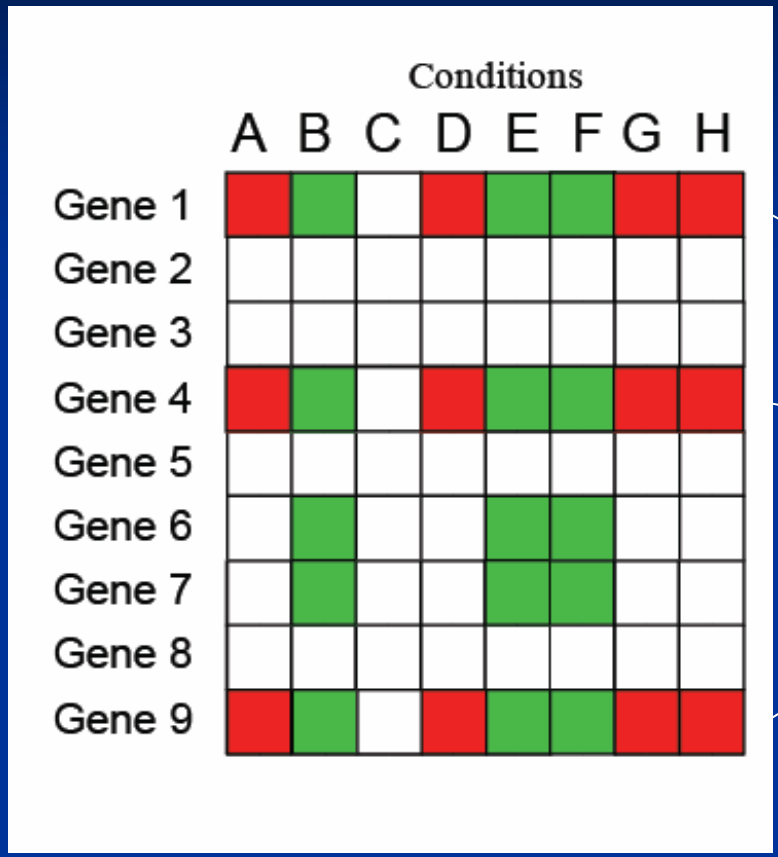
b) Alto verso il basso

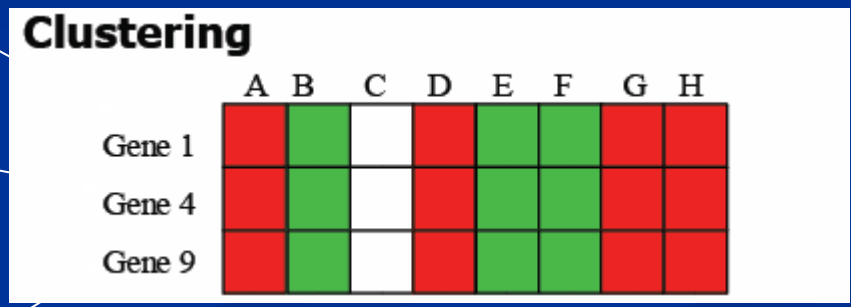


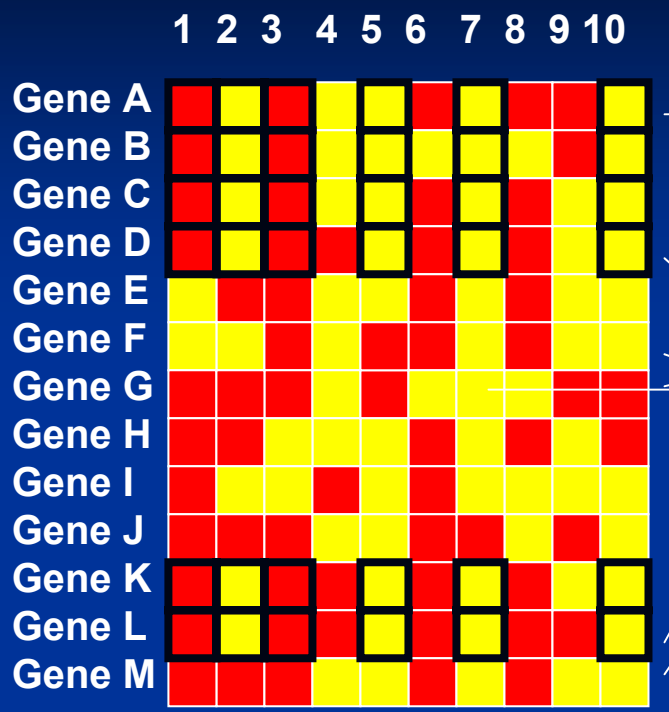
Lo scopo, nell'analizzare la matrice, è quello di poter identificare gruppi di geni con profili di espressione simili. Spesso i geni appartenenti allo stesso cluster sono detti coespressi.
La ragione per cui si cercano geni coespressi sono:

1) si è evidenziato che geni funzionalmente correlati sono coespressi

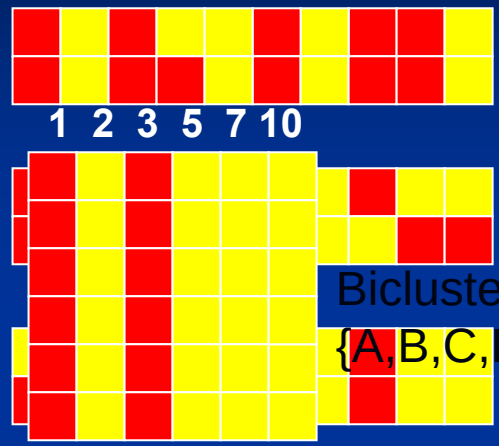
2) geni coespressi possono dare informazioni sui meccanismi regolatori


Conditions
A B C D E F G H
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9


Clustering
A B C D E F G H
Gene 1
Gene 4
Gene 9



Clustering…



Bicluster {1,2,3,5,7,10}
{A,B,C,D,E,F}

Non esistono livelli di espressione simili considerando tutte le condizioni…

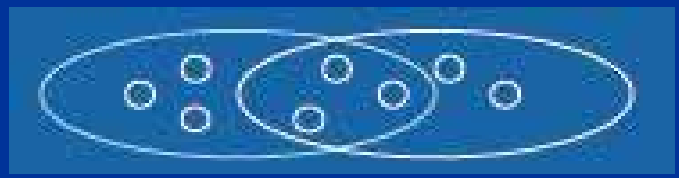
Soluzione: Cluster simultaneo di righe e colonne (**Biclustering**)

BICLUSTERING

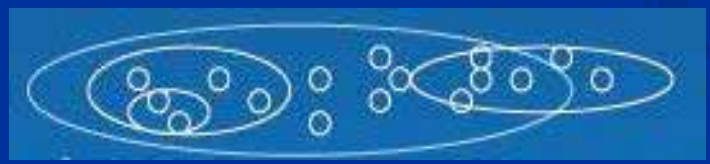
Tecnica di data mining che raggruppa i geni rispetto a sottoinsiemi di condizioni.

Ciò ci permette di rappresentare meglio:

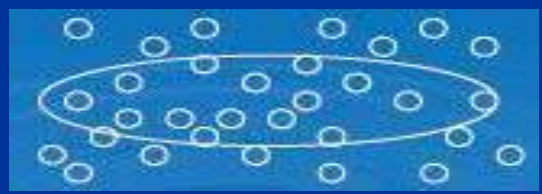
- **Cluster correlati**



- **Segnali locali**



- **Rumore**



Bicluster = sottomatrice della matrice di espressione genica che soddisfa alcune caratteristiche di omogeneità

Bicluster(I,J)= sottomatrice kxs

$I = \{i_1,\ldots\ldots,i_k\}$ $J = \{j_1,\ldots\ldots,j_s\}$

Può essere di vari modi:

- **valori costanti**
- **valori costanti sulle righe o sulle colonne**
- **valori coerenti**

Algoritmo di Cheng e Church(node deletion)

Similarità = coerenza di geni e condizioni nel subset

mean squared residue score (H)

$$R(a_{ij}) = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}$$

a_{iJ} **media dei valori nella subm. lungo la linea i**

a_{Ij} **media dei valori nella subm. lungo la colonna j**

a_{IJ} **media di tutti i valori nella submatrice**

$$H_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} R^2$$

- **Matrice costante: 0**
- **Matrice con singolo elemento: 0**
- **Matrice con elementi scelti random in un intervallo [a,b] , H = $(b-a)^2/12$**

H elevato ⇒ i dati non sono correlati

H basso ⇒ i dati sono correlati

Media-matrice = 6.5

1	2	3	2	
4	5	6	5	media r.
7	8	9	8	
10	11	12	11	
5.4	6.4	7.4		

media c.

R(1)=1-2-5.4+6.5=0.1

R(2)=2-2-6.4+6.5=0.1

……..

R(12)=12-11-7.4+6.5=0.1

$H=(0.01 \times 12)/12 = 0.01$

Se i valori fossero stati scelti in modo casuale tra [1,12]

$H = (12-1)^2/12 = 10.08$

OBIETTIVO = data una matrice di espressione A e un valore δ, cercare un bicluster con massimo un H=δ

- **Input: A e δ >=0 minimo H accettabile**
- **Output: A(I,J), detto δbicluster (H>=δ)**
- **Inizializzazione: A(I,J)=A**
- **Iterazione: si calcola a_{iJ} per ogni i εI , a_{Ij} per ogni j di J, a_{IJ} e H(I,J).**

  se H<=δ si ritorna A(I,J)

  si cerca la riga con valore più alto $d(i) = \frac{1}{|J|}\sum_{j \in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$

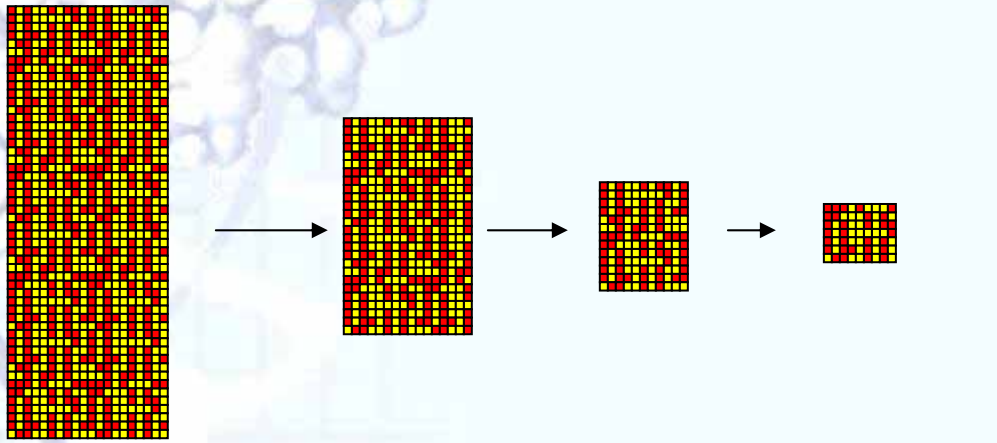
  e si cerca la colonna con valore più alto $d(j) = \frac{1}{|H|}\sum_{i \in I}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$

  si elimina la colonna o la riga aggiornando I o J

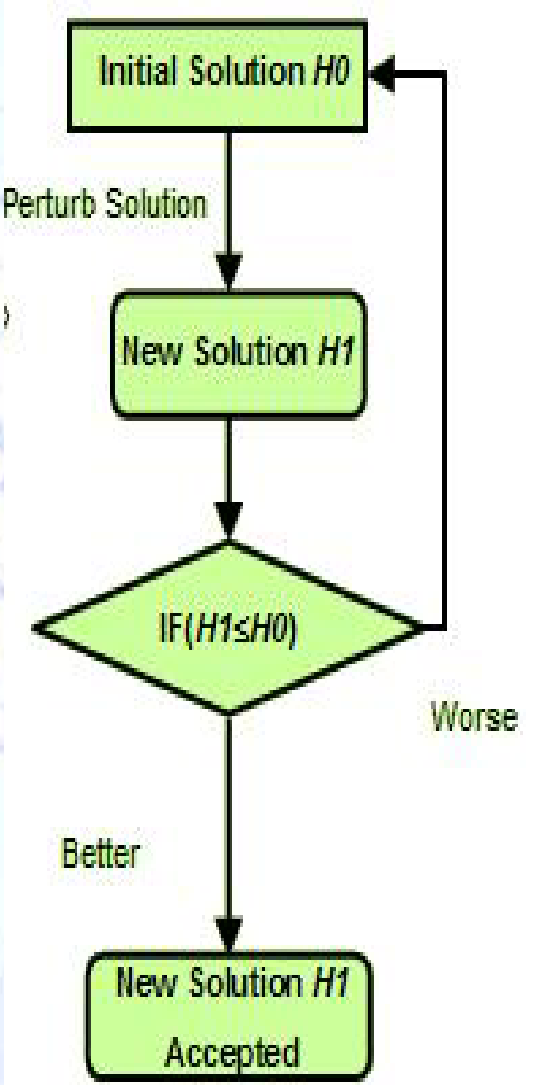
Input:

Matrice di dati

δ = 300



Score: 1,052 **Score: 543** **Score:423** **Score: 300**



Ricerca Locale(s);

s*=s; fine=false

repeat

s:=migliore soluzione I(s)

if f(s)<f(s*) then

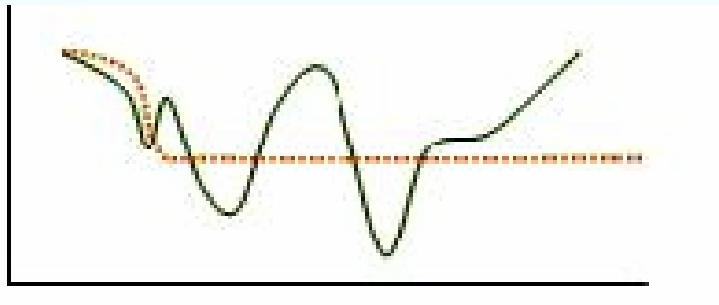
s*:=s;

else fine=true;

until not fine;

return s*;

end



SIMULATED ANNEALING

Nato per simulare l'annealing dei solidi

Processo termico atto ad ottenere stati di bassa energia in un solido immerso in un bagno di calore.

1) **Si aumenta la temperatura ad un valore tale che il solido si scioglie.**

2) **Si decresce lentamente la temperatura fino a quando la particelle si sistemano in uno stato base in cui l'energia è minima.**

Il sistema si dice in equilibrio termico alla temperatura T se la probabilità p(Ei) di uno stato avente energia Ei è governata dalla distribuzione di Boltzmann.

$$p(Ei) = g_i \frac{\exp\left(-\frac{Ei}{K_B T}\right)}{\sum_J \exp\left(-\frac{Ej}{K_B T}\right)}$$

ALGORITMO DI METROPOLIS

Algoritmo che simula l'evoluzione del solido verso l'equilibrio termico.

- **Dato lo stato corrente i del solido con energia Ei**
- **Perturbandolo si genera un nuovo stato j con Ej**
- **Se Ei-Ej>=0 lo stato j diviene il nuovo stato corrente**
- **Altrimenti lo stato j diviene il nuovo stato corrente con probabilità**

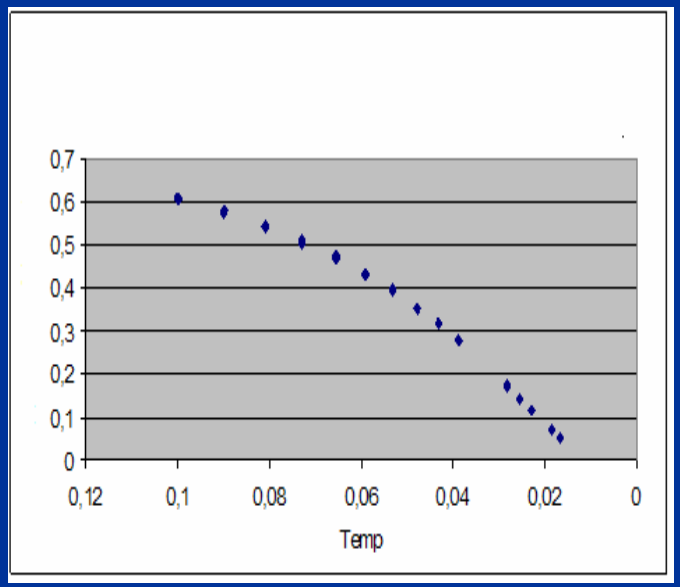
$$e^{\left(\frac{Ei-Ej}{kT}\right)}$$

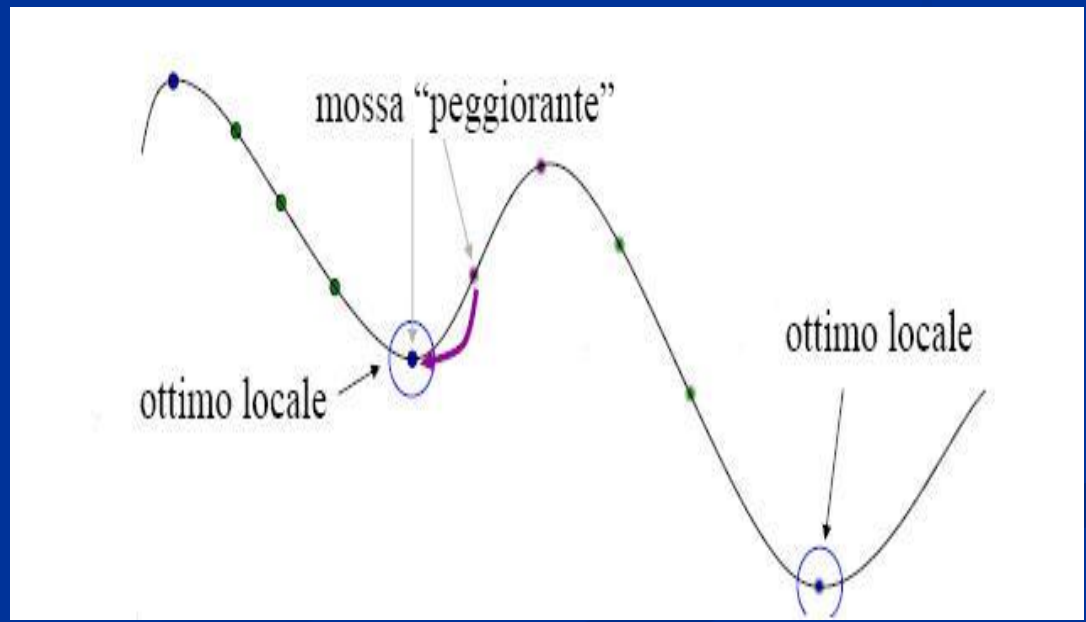
T = temperatura del bagno di calore

K = costante di Boltzmann

All'inizio la ricerca salta da un punto all'altro individuando le direzioni o le aree in cui è più probabile trovare l'ottimo globale.

A basse temperature le soluzioni vengono localizzate in un dominio promettente.





Simulated annealing → problemi di ottimizzazione

- **soluzione di un problema = stati di un sistema fisico**
- **costo di una soluzione = energia di uno stato**

  funzione fitness che definisce se una perturbazione determina un miglioramento
- **parametro di controllo = temperatura**

To ?

velocità di raffreddamento

Elevata = blocco in un minimo

Bassa = aumentano gli stati, quindi il costo

La temperatura viene ridotta ad ogni stato moltiplicandola per una costante chiamata

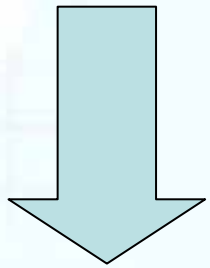
0<cooling ratio>1

Metodo logaritmico

- **La procedura che genera nuove soluzioni deve essere il più veloce possibile.**

Criterio di interruzione

- **numero di soluzioni trovate è < di un valore fissato**
- **valore finale di temperatura**



Stabiliti in base al problema che si sta affrontando

Parametri iniziali:

- **Funzione fitness = mean squared residue score**
- **To = definita tramite esperimenti**
- **$T(i) = T(i-1)/(1+\sigma)$**

  $\sigma = 0.1$ cooling ratio

criteri di arresto:

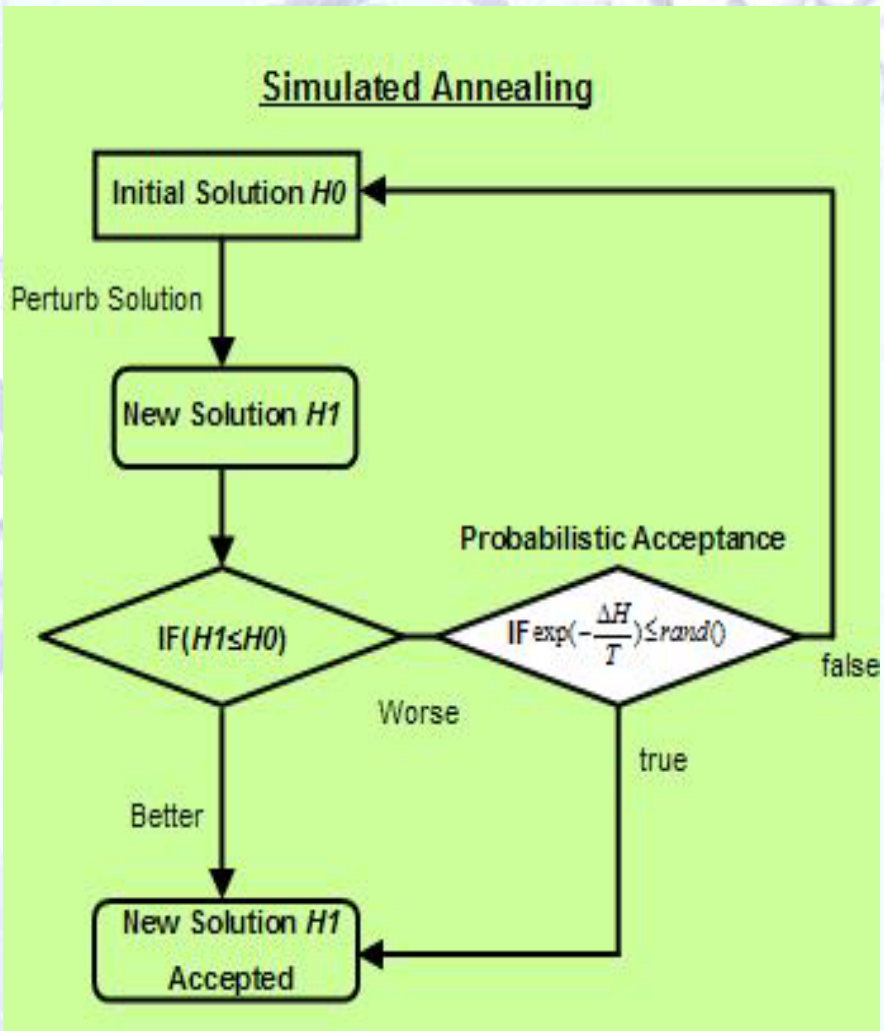
- **$\delta = 300, 200, 100$ → node deletion (CCND/2)**
- **bicluster minimo 10x10 → node deletion modificato (CCND2)**

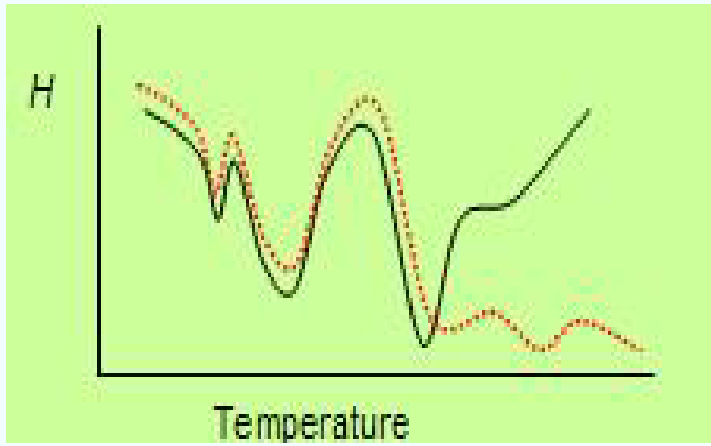
SAB(f, xo, to, r, a, s, M, l, δ)

$x \leftarrow x_0$
$t \leftarrow t_0$
While$(t > t_{Min})$
  While$(a_{count} < a$ AND $s_{count} < a)$
   $x_{new} \leftarrow$ GenerateNewSolution(M, x)
  if$f(x_{new}) < f(x)$
    then $x \leftarrow x_{new}$
  else if $\exp(-\frac{\Delta E}{T})>$ random(0,1)
    then $x \leftarrow x_{new}$
  if $(f(x_{new}) \leq \delta)$ AND SizeOf$(x_{new})>$SizeOf(x)
    then $l \leftarrow$ SizeOf(x_{new})
   $t \leftarrow$ cool$(t, rate)$
$x \leftarrow$ NodeAdditon(x)
return x

Funzione genera(M,x)

$row_x \leftarrow$number of rows in x
$col_x \leftarrow$number of rows in x
if($row_x \geq col_x$)
then generate random r, range$[0, \frac{row_x}{col_x}] \in \mathbb{R}$
if ($r = 0$)
then flip random column in M to generate X_{new}
else
then flip random row in M to generate X_{new}
else if($row_x < col_x$)
then generate random r, range$[0, \frac{col_x}{row_x}] \in \mathbb{R}$
if ($r = 0$)
then flip random row in M to generate X_{new}
else
then flip random column in M to generate X_{new}
return x_{new}

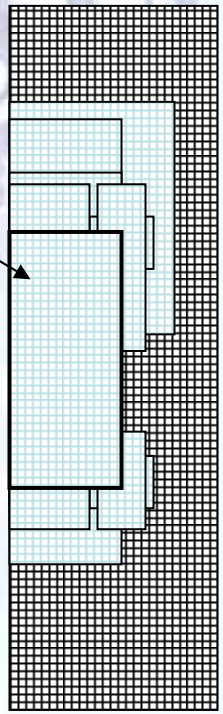

Simulated Annealing
Initial Solution H0
Perturb Solution
New Solution H1
Probabilistic Acceptance
IF(H1≤H0)
IF exp(−ΔH/T)≤rand()
false
Worse
true
Better
New Solution H1
Accepted


H
Temperature

Simulated Annealing

Per evitare di cercare lo stesso bicluster questo viene sostituito con numeri random

Colonne (Condizioni)



Righe (Geni)

Mean Squared Residue Score

1

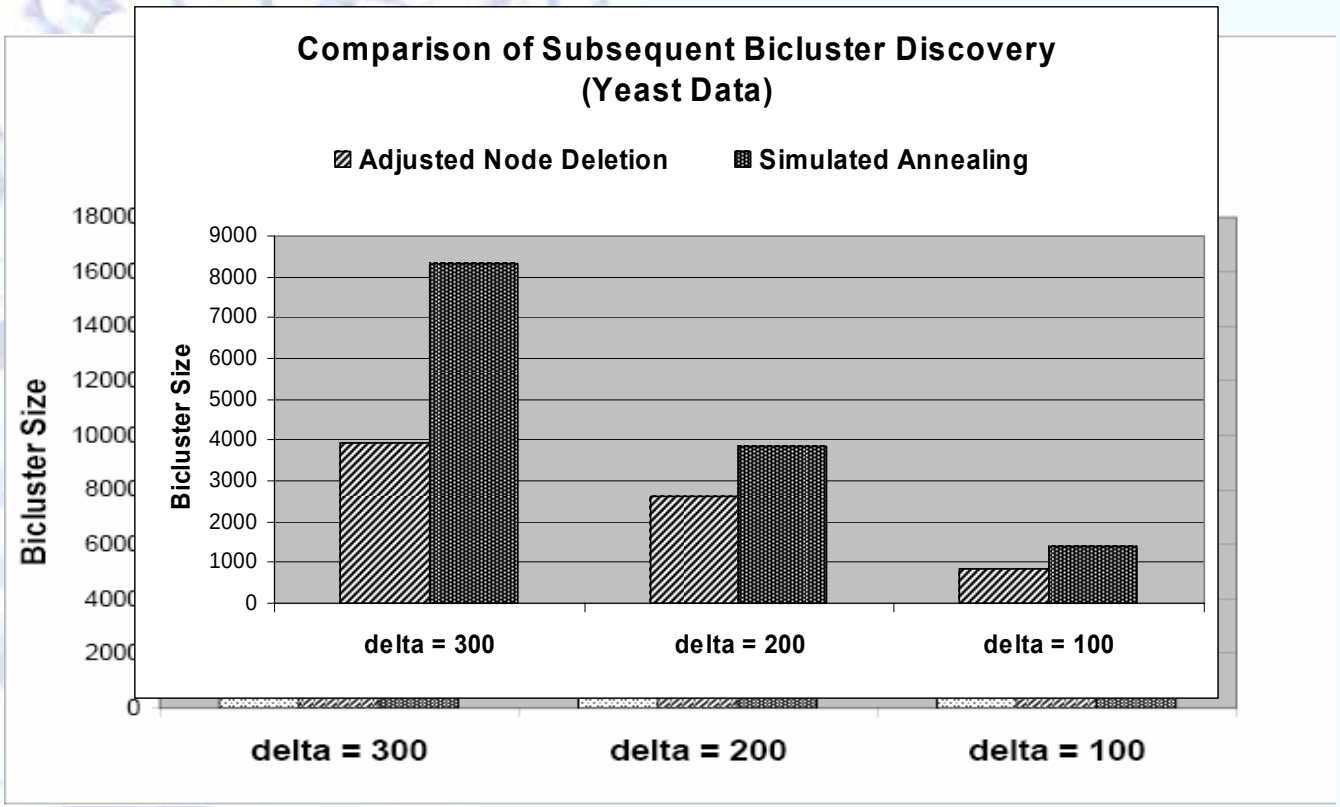
Bicluster massimo

Delta si blocca e la dimensione aumenta

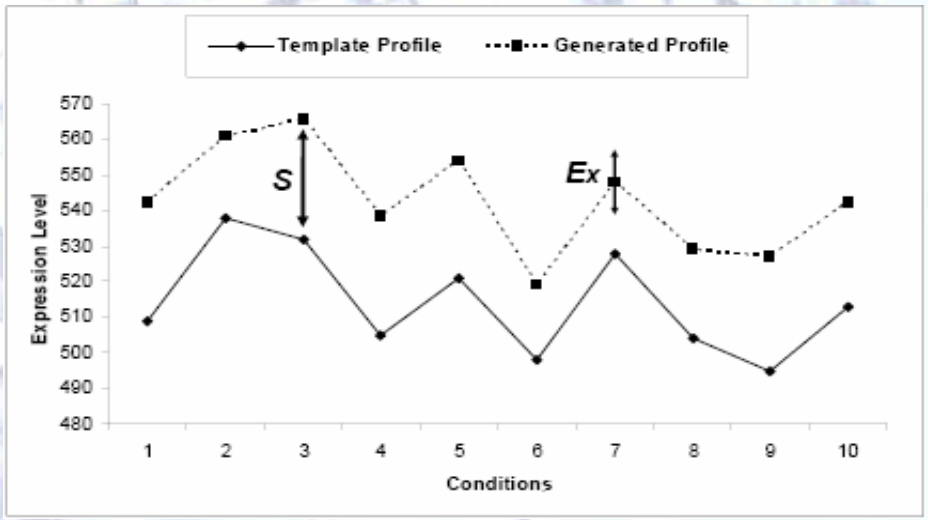
Dati

geni **condizioni**

2884 **17** **lievito**



Partendo dal profilo di espressione iniziale si effettua uno shift random



Ex = σ(x).ε.rx

σ(x)= deviazione standard del nuovo profilo genico

ε= costante [0,1] (livello errore e qualità del bicluster)

rx= variabile random [-1,1] (il livello di espressione sia più alto o iù basso)

$$X = \{x_1,\ldots\ldots,x_n\}$$

$$Y = \{x_1+(S+E_1),\ldots\ldots,x_n+(S+E_n)\}$$

2774 geni 27 condizioni sclerodermia

3051 geni 38 condizioni linfoma

	Node Deletion	Adjusted Node Deletion	Simulated Annealing	Node Deletion	Adjusted Node Deletion	Simulated Annealing
Yeast	First Bicluster			Second Bicluster		
$\delta = 300$	15165	15750	**16460**	*9012*	3930	**8320**
200	8463	9540	**10360**	*4972*	2630	**3860**
100	2520	2700	**2940**	*1260*	830	**1390**
Scleroderma						
$\delta = 300$	13590	**18260**	**18230**	*4320*	**6780**	6310
200	7296	12920	**13210**	*7876*	3290	**4030**
100	2730	**5170**	**5140**	*1570*	830	**850**
Lymphoma						
$\delta = 300$	1344	**3320**	3220	*518*	1740	**1810**
200	1032	**2510**	2460	*300*	**1370**	1200
100	851	**1780**	**1790**	*136*	**1050**	810

Cercare con SAB biclusters che abbiano un autentico significato biologico.

550 geni su 2884 sono annotati nella banca dati KEGG.

Bic.	# Genes	Dominant FM	Genes in FM	P-value
1	81	Ribosomal Proteins(96)	61	2.15×10^{-38}
		Glycolysis/Glucogenisis(26)	5	0.16
2	59	Basal Transcription Factors(10)	6	1.6×10^{-4}
		Nucleotide Metabolism(81)	16	4.02×10^{-3}

P-value = probabilità che (x geni / totale dei geni) appartenga alla funzione y

Più è basso e più è significativo

Bicluster 1 contiene 75% di geni che codificano per proteine ribosomali

Bicluster 2 contiene 40% di geni che codificano per fattori di trascrizione.

Il SAB trova biclusters significativi ma, in questi non domina totalmente un gruppo funzionale

- **dati incompleti**
- **la ricerca top down trova biclusters che non rispecchiano la situazione '*in vivo*'**

FUTURO: SAB ⟶ bottom up