Clustering

Il clustering organizza i geni in gruppi (cluster) con simili pattern di espressione.
Spesso i geni apparteni allo stesso cluster sono detti coespressi.
Le ragioni per cui si cercano geni coespressi sono:

1. Ci sono evidenze che molti geni funzionalmente correlati sono coespressi. Ad esempio geni codificanti per elementi di un complesso proteico solitamente hanno simili pattern di espressione.
2. Geni coespressi possono dare informazioni sui meccanismi regolatori. Se un sistema regolativo controlla due o più geni questi risulteranno essere coespressi.

Differenti tecniche di clustering sono state sviluppate ed applicate all'analisi di dati da microarray.
Il clustering raggruppa i geni che mostrano simile comportamento tra i differenti punti sperimentali, non tra le repliche.
Quindi all'interno degli algoritmi di clustering bisogna inserire un unico dato per ogni gene per ogni gruppo di replicati.
Bisogna inoltre tenere presente che tale tipo di indagine va condotta esclusivamente nel caso in cui i punti sperimentali indagati siano superiori a due (un time-course ad esempio).
Un clustering condotto su un esperimento con due punti è completamente inutile.

Distance metrics

Il principio alla base del clustering è quello delle distance metrics.
Queste, nel nostro caso rappresentano una distanza di espressione di ogni gene rispetto agli altri.
In base a questo principio vengono raggruppati nello stesso cluster tutti quei geni che mostrano distanza minima tra di loro sono quindi vicini biologicamente parlando.

Dicotomia

Anche per quel che riguarda il clustering esistono numerosi tipi di algoritmi differenti e abbiamo solo l'imbarazzo della scelta. I 2 tipi di clustering più utilizzati sono il GERARCHICO e il K-MEANS. Nella classificazione dei tipi di clustering esiste inoltre una certa dicotomia.
Distinguiamo tra:

- GERARCHICO
- NON GERARCHICO

- SUPERVISIONATO
- NON SUPERVISIONATO

- AGGLOMERATIVO
- DIVISIVO

Che significa?

- Il **CLUSTERING GERARCHICO** è il più semplice da visualizzare, per far questo vengono utilizzati dei dendrogrammi che oltre a formare i gruppi mettono anche in relazione i singoli geni ed esperimenti a partire da una radice che contiene tutta la gerarchia e che si va a poco a poco ramificando fino ad arrivare al singolo elemento.

Questo tipo di clustering può essere **SUPERVISIONATO** cioè utilizzare informazioni note sui geni per guidare l'algoritmo.
Ed è **AGGLOMERATIVO** cioè parte da un singolo gene a caso ed inizia a costruire i gruppi gene per gene.

- Il **CLUSTERING K-MEANS** utilizza algoritmi computazionalmente più impegnativi, non è visualizzabile con dendrogrammi e generalmente non è supervisionato.
  Fa parte della categoria di clustering divisivi in quanto parte dal totale degli elementi per dividerli successivamente in gruppi.
  Il numero di gruppi finale in cui dividere gli elementi deve essere impostato a priori dall'utente.

- **SELF ORGANIZING MAP (SOM)** divisivo basato su reti neurali quindi supervisionato.

- **PRINCIPAL COMPONENT ANALISYS (PCA)** basato su algoritmi matematici che calcolano un' espressione media per geni simili prima di inserirli nei cluster.

Ci sono tanti altri algoritmi e loro variazioni per il clustering, l'importante per adesso è capirne solo i principi.
Bisogna però capire bene un'ultima cosa: è indispensabile eseguire clustering solo dopo aver filtrato statisticamente i risultati; costruendo una serie di gruppi attorno ai geni dobbiamo evitare di portarci dietro il più possibile geni con valori non affidabili per diminuire il rischio di costruire cluster errati.

Vediamo adesso di approfondire un po' il tutto.
L'insieme dei dati da inserire nel clustering rappresenta una matrice di espressione.

Matrice di espressione

Una serie di *m* array misurano i livelli di espressione espressione in *m* differenti condizioni sperimentali.
Denotiamo la matrice X di misura (*n*-geni x *m*-array) come tavola dei dati di espressione.
Qui x_{ij} è il $\log_2$ del rapporto di espressione del gene *i* nel campione *j*.
Quindi il vettore nella riga *i* della matrice X indica i rapporti di fluorescenza del gene *i* in tutti i differenti campioni.
Mentre il vettore nella colonna *j* della matrice X indica i rapporti di fluorescenza di tutti i geni nel campione *j*.

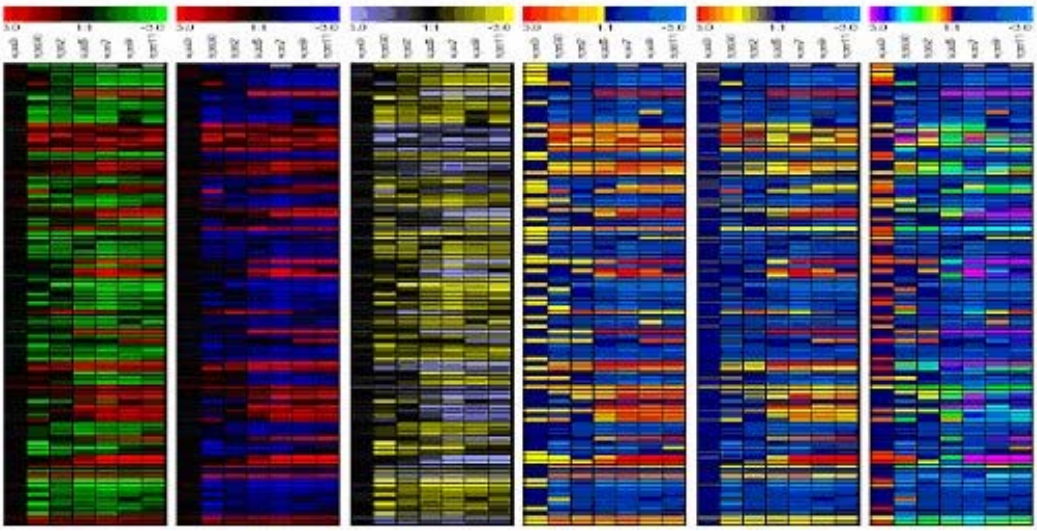
$$x_{ij} = \log_2 \frac{Sample1ij}{Sample2ij}$$

x_{ij} è negativo se $Sample2_{ij} > Sample1_{ij}$,
positivo se $Sample1_{ij} > Sample2_{ij}$,
zero se $Sample2_{ij} = Sample1_{ij}$.

Quindi dette matrici non saranno altro che tabelle in cui in ogni riga avremo i valori di uno stesso gene, mentre in ogni colonna avremo i valori di uno stesso array.
Nella matematica odierna, per vettore si intende più in generale un insieme ordinato di quantità dette componenti, nel nostro caso ogni vettore è rappresentato da un singolo gene le cui componenti sono i segnali di espressione.
Praticamente ogni gene è rappresentato da un vettore all'interno della matrice ed il principio alla base del clustering non fa altro che associare vettori che mostrano componenti paragonabili (simili segnali di espressione)

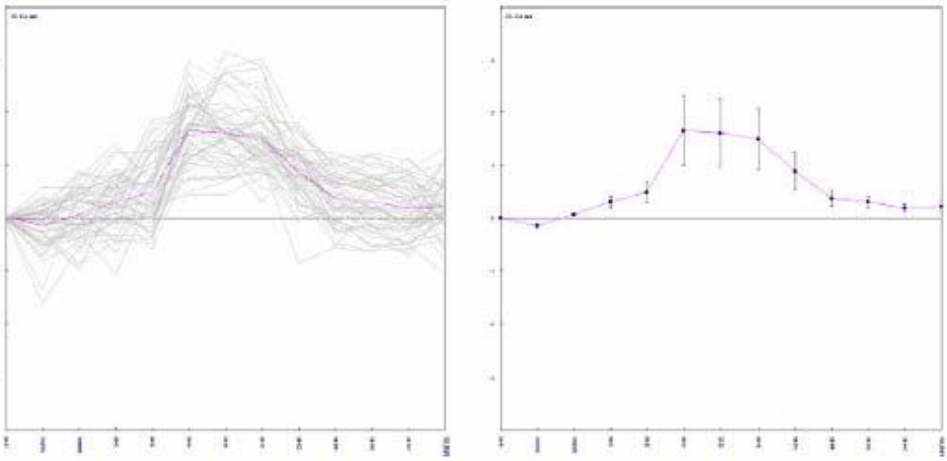
Rappresentazione grafica

Poiché una grossa lista di numeri è difficile da valutare, i dati grezzi sono convertiti in rappresentazioni grafiche rappresentando ogni dato con un colore che lo caratterizza quantitativamente e qualitativamnte. Ciò permette al biologo una esplorazione e comprensione del risultato più intituitiva.
Generalmente i colori utilizzati vanno dal verde saturato (valore max negativo) al rosso negativo (valore max positivo), i geni il cui ratio è zero sono solitamente neri. Ma sono possibili anche altri tipi di rappresentazioni.



I cluster possono essere visualizzati anche in altre maniere:

1. Ogni gene in un cluster viene plottato separatamente. EXPRESSION GRAPHS
2. Vengono plottate la media e la deviazione standard di tutti i geni del cluster. CENTROID GRAPHS



Rifinitura dati

Ci sono differenti procedure utilizzate prima di eseguire il clustering su una serie di dati scelti.
Solitamente si usa immettere nei programmi di clustering i rapporti trasformati in $\log_2$.
Si può eseguire il centraggio sulla media o sulla mediana o altro.

Distance metrics (Similarity distances)

Tutti gli algoritmi di clustering usano misurare la somiglianza tra vettori per comparare i pattern di espressione.
La distanza tra due geni e/o esperimenti è computata sommando le distanze tra i loro rispettivi vettori.
Come sono calcolati questi valori dipende dalla Similarity distances utilizzata nella determinazione della matrice delle distanze.

. Pearson correlation coefficient
. Uncentered Pearson correlation coefficient
. Squared Pearson correlation coefficient
. Averaged dot product
. Cosine correlation coefficient
. Covariance
. Euclidian distance
. Manhattan distance
. Mutual information
. Spearman Rank-Order correlation
. Kendall's Tau.

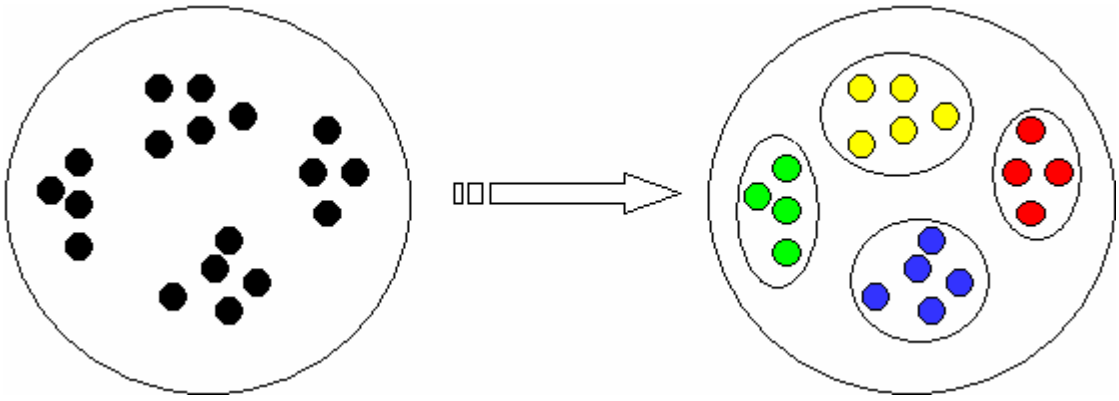
Quindi ogni distanza misurata non sarà altro che una quantizzazione della relazione lineare tra due serie di misure x e y.

Principi

Ci sono principalmente due modi in cui le matrici di espressione genica risultanti da un esperimento di microarray possono essere studiate:

1. Comparando i profili di espressione dei singoli geni nella matrice e
2. Comparando i profili di espressione degli interi esperimenti nella matrice.

Inoltre è anche possibile una combninazione dei due.
Se due geni sono espressi in maniera simile possiamo ipotizzare che sono coregolati o comunque funzionalmente correlati.
Il clustering può essre definito come il processo di raggruppamento di oggetti in gruppi diversi in base alla loro similarità.



Clustering gerarchico - HCL

Una procedura non supervisionata che trasforma le distance matrix in una gerarchia di gruppi.
La gerarchia è rappresentata tramite dendrogrammi in cui ogni cluster si trova comunque inserito all'interno di altri cluster più grandi.

Algoritmo

1. Calcola le distanze tra tutti gli oggetti e costruisce la matrice. Ogni gene rappresenta un cluster contenente solo se stesso.
2. Trova i 2 cluster *r* e *s* con la minima distanza tra loro.
3. Fonde i due cluster *r* ed *s* e rimpiazza *r* con il nuovo cluster.
4. Elimina *s* e ricalcola le distanze che sono state interessate dalla fusione.
5. Ripete le fasi 2, 3 e 4 finchè il numero totale dei cluster non diviene 1, cioè finchè non sono stati presi in considerazione tutti i geni.

Parametri

Linkage Method

Questo parametro è utilizzato per indicare quale algoritmo utilizzare per calcolare le distanze tra due cluster quando si costruisce il dendrogramma. Si distinguono

Single Linkage:
le distanze sono misurate da ogni membro di un cluster ad ogni membro dell'altro cluster. Tra tutte queste distanze la *minima* è considerata la distanza tra i cluster.

Average Linkage:
la misuara della distanza tra due cluster è considerata la media della distanza di ogni membro del cluster da ogni membro dell'altro.

Complete Linkage:
le distanze sono misurate da ogni membro di un cluster ad ogni membro dell'altro cluster. Tra tutte queste distanze la *massima* è considerata la distanza tra i cluster.

Support Trees (resampling HCL) – ST

Questa evoluzione del clustering gerarchico mostra gli alberi ottenuti dal precedente, ma in più calcola un 'supporto statistico' per i nodi del dendrogramma basato sul rimescolamento dei dati.
Il clustering gerarchico è un tipo di clustering agglomerativo e quindi nel costruire i gruppi l'algoritmo parte da un gene a caso e a poco a poco inizia a costruirci i cluster sopra.
Il risultato finale quindi può essere dipendente dal gene che per caso viene scelto per primo.
Il resampling non fa altro che ripetere la stessa operazione di clustering sullo stesso set di dati, ma partendo da un gene diverso.
Il software in questo modo può valutare quante volte un gruppo di geni viene clusterizzato assieme, una misura questa dell'attendibilità del cluster formato.
Qui c'è da indicare che algoritmo utilizzare per il resampling.

K-Means – KMC

Questo tipo di clustering è utile quando l'utente ha già una ipotesi sul numero dicluster che i geni dovrebbero formare. K è infatti il numero di cluster da formare con il set di dati e va impostato a priori dall'utente.

Algoritmo

1. Assegna tutti i geni di un esperimento ad uno dei *k* cluster.
2. Calcola la media o la mediana per ognuno dei cluster.
3. Calcola la distanza tra ogni oggetto e la media o la mediana di ogni cluster.
4. Sposta ogni oggetto nel cluster la cui media è più vicina a quell'oggetto.
5. Ricalcola media dei cluster interessati dalla riallocazione.
6. Ripete le operazioni 3,4 e 5 finchè non sono necessari più spostamenti o ha raggiunto il massimo di iterazioni impostate.

Self Organizing Map - SOM

Basato su reti neurali: l'algoritmo impara dai dati e costruisce i gruppi in base a quello che ha imparato.
Una SOM è una griglia rettangolare od esagonale in cui ogni cluster è rappresentato da un elemento.

Algoritmo

1. Fase di training in cui si prende un gene a caso e si misurano le misure di distanza di questo da tutti i nodi della mappa. Si passa ai geni successivi modificando ogni volta i nodi della griglia.
2. Fase di clustering in cui lo spazio virtuale dell'esperimento viene ridotto e sovrapposto alla griglia ed ogni gene associato ad un vettore di essa.

Parametri

Dimension X ed Y:

rappresentano le dimensioni della risultante topologia della SOM. Indicano quindi anche il gruppo di cluster da costruire (es. da X = 3 e Y = 2 risulteranno 6 cluster)

Cluster Affinity Search Tecnique – CAST

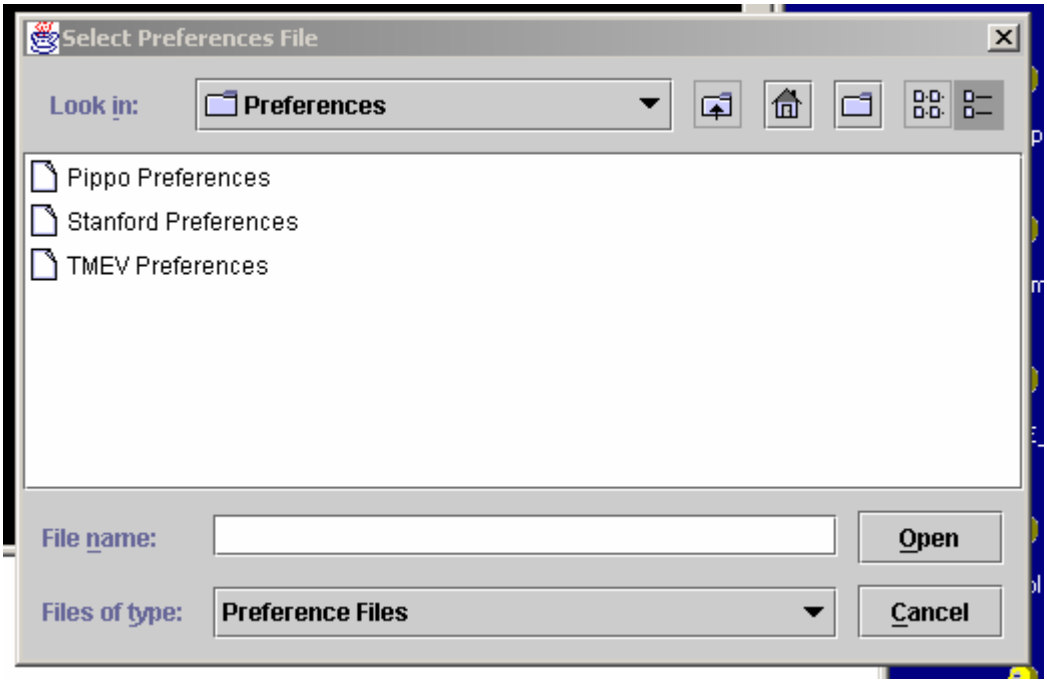
L'utente deve impostare un valore tra 0 e 1 come soglia di affinità (il reciproco della distance metric tra due geni) che non dovrà essere superata da alcun gene nel cluster.
L'algoritmo lavora costruendo cluster e aggiungendo geni a questi finchè è rispettato il threshold.
Maggiore è il valore soglia più stringente sarà la nostra analisi e più cluster verranno prodotti.

Principal Component Analisys

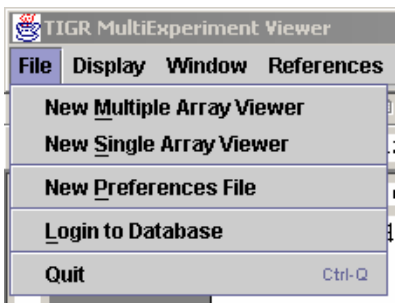
PCA sono utilizzati per attribuire la grossa variabilità dei dati ad un ridotto set di variabili chiamate principal component. Questo elimina il rumore di fondo del dataset e concentra l'investigazione verso gli aspetti che introducono la maggior parte della variabilità nei dati.
Una certa frazione della variabilità totale dei dati è assegnata ad ogni principal component.
Questo classifica i component in ordine di variabilità decrescente.
Le prime tre principal component sono utilizzate per mappare ogni gene nello spazio tridimensionale.

TIGR MeV – TmeV

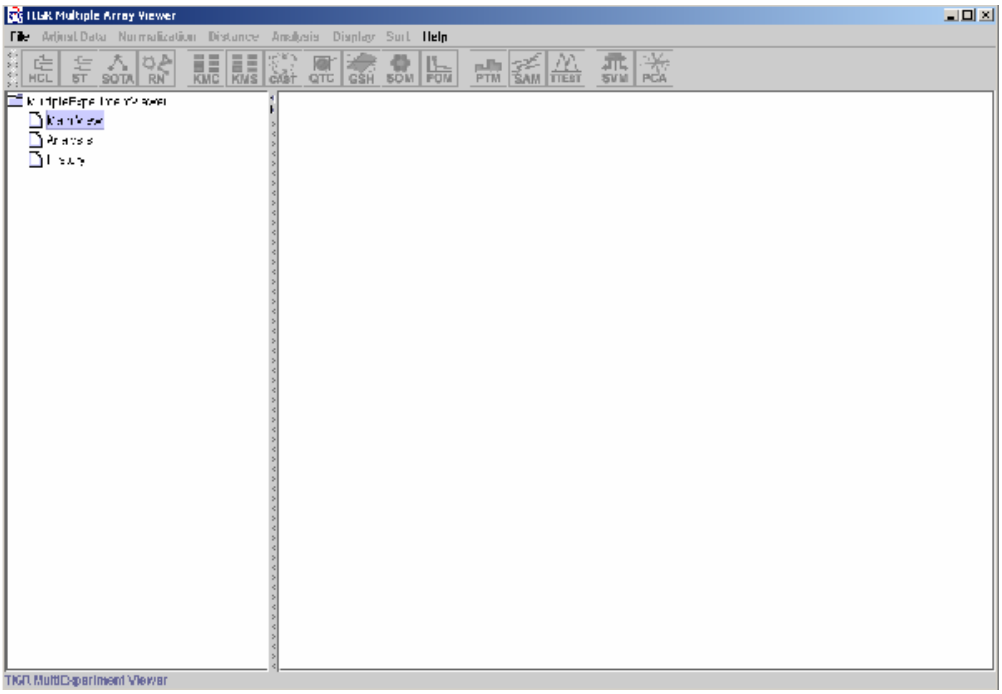
TmeV è un programma open-source sviluppato al TIGR da Quackenbush ed il suo gruppo per il clustering dei dati da microarray.
Il programma è scritto in java ed è ormai giunto alla release 2.1 quindi funziona abbastanza bene!
E' liberamente scaricabile dal sito del TIGR (www.tigr.org) e per esser utilizzato richiede la preinstallazione delle librerie java standard e di quelle 3D.
All'apertura viene chiesto di selezionare un file di preferenze che serve per indicare al software il formato dei file di espressione da leggere



Selezioniamo il tipo Stanford.
A questo punto selezioniamo File -> New Multiple Array Viewer



Si apre così la finestra del programma vero e proprio da cui si può fare tutto ciò di cui abbiamo parlato fino ad ora ed anche di più!



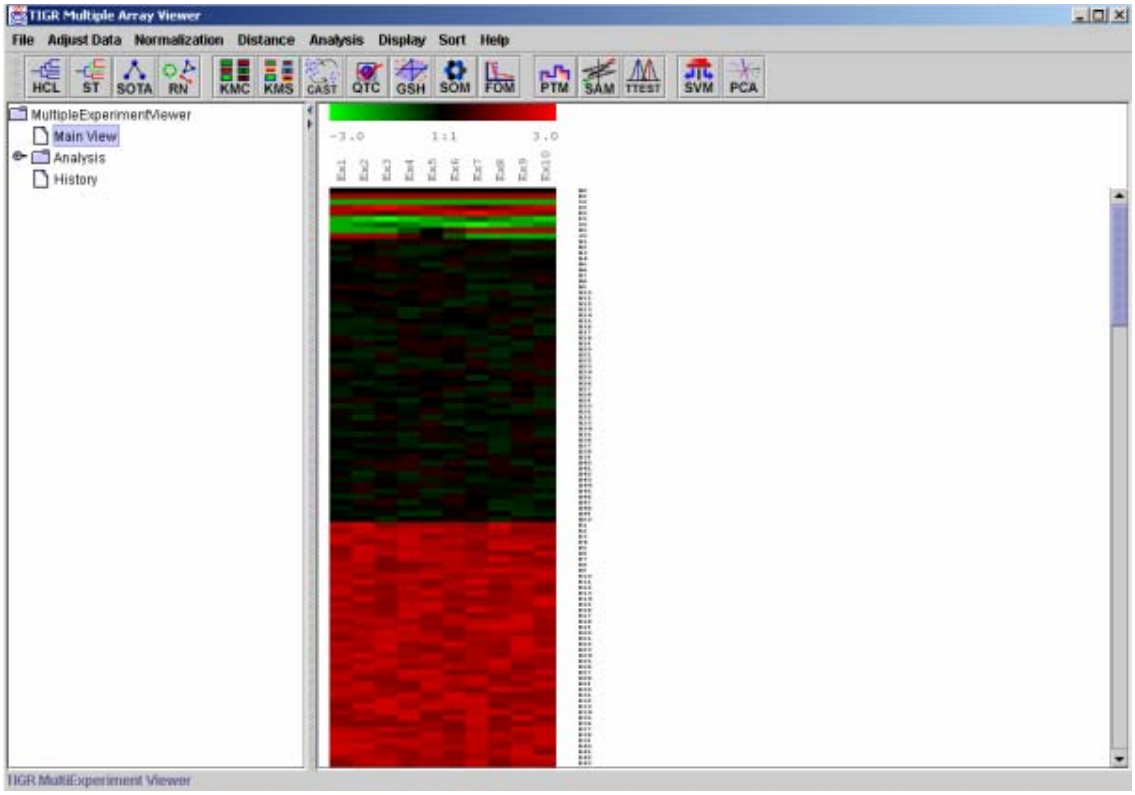
Selzionare File -> Add Expreriment From Stanford File ed inserire il file dei risultati Affymetrix che avrete opportunamente preparato per renderlo compatibile per le specifiche Stanford.
Il file inoltre dovrà contenere oltre ai normali tag Stanford, i ratio per ogni comparazione invece dei segnali assoluti dai singoli esperimenti.
Quello che si va a clusterizzare sono i rapporti e non i segnali.

	A	B	C	D	E	F	G	H
1	YORF	NAME	GWEIGHT	Ex1	Ex2	Ex3	Ex4	Ex5
2	EWEIGHT			1	1	1	1	1
3	A0	A0	1	0	0	0	0	0
4	B0	B0	1	2	2	2	2	2
5	C0	C0	1	-2	-2	-2	-2	-2
6	D0	D0	1	2	2.5	3	2.5	2
7	E0	E0	1	2	1.5	1	1.5	2
8	F0	F0	1	-2	-2.5	-3	-2.5	-2
9	G0	G0	1	-2	-1.5	-1	-1.5	-2
10	H0	H0	1	-2	-2	-2	-1	0
11	J0	J0	1	2	2	2	1	0
12	A1	A1	1	0.246456878	0.22613566	-0.279917733	0.278944897	0.416614486
13	A2	A2	1	-0.323161859	0.007695167	0.354104906	-0.324017974	-0.408117939
14	A3	A3	1	0.202584754	-0.086175734	-0.329190141	-0.38246136	-0.297690969
15	A4	A4	1	-0.365973455	0.253069289	0.369742468	-0.145891188	0.032014186

Il formato Stanford prevede:

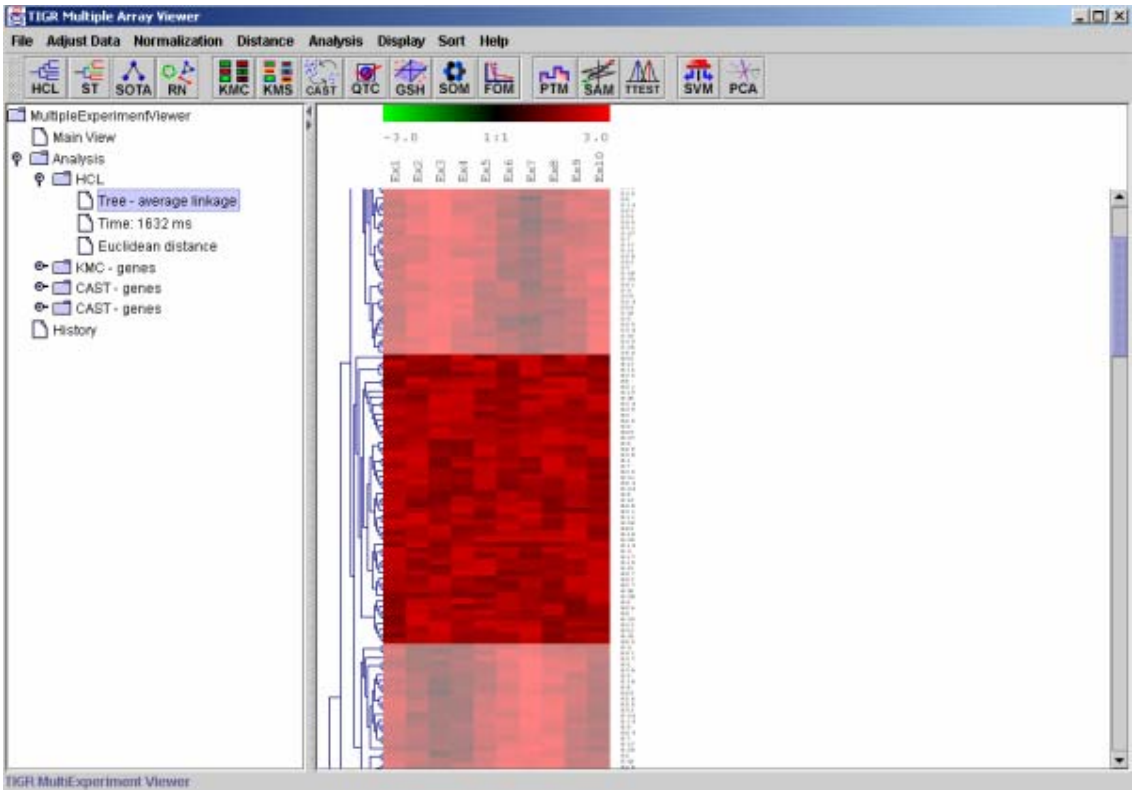
- Nella prima riga le intestazioni.
- Nella seconda riga un peso assegnato ad ogni esperimento (porre tutti i valori a 1).
- Nelle prime due colonne l' ID e le descrizioni dei geni.
- Nella terza colonna un peso associato ad ogni gene (porre tutti i valori a 1).

A questo punto l'esperimento è caricato nella main view che fa vedere i dati inseriti in maniera grafica.

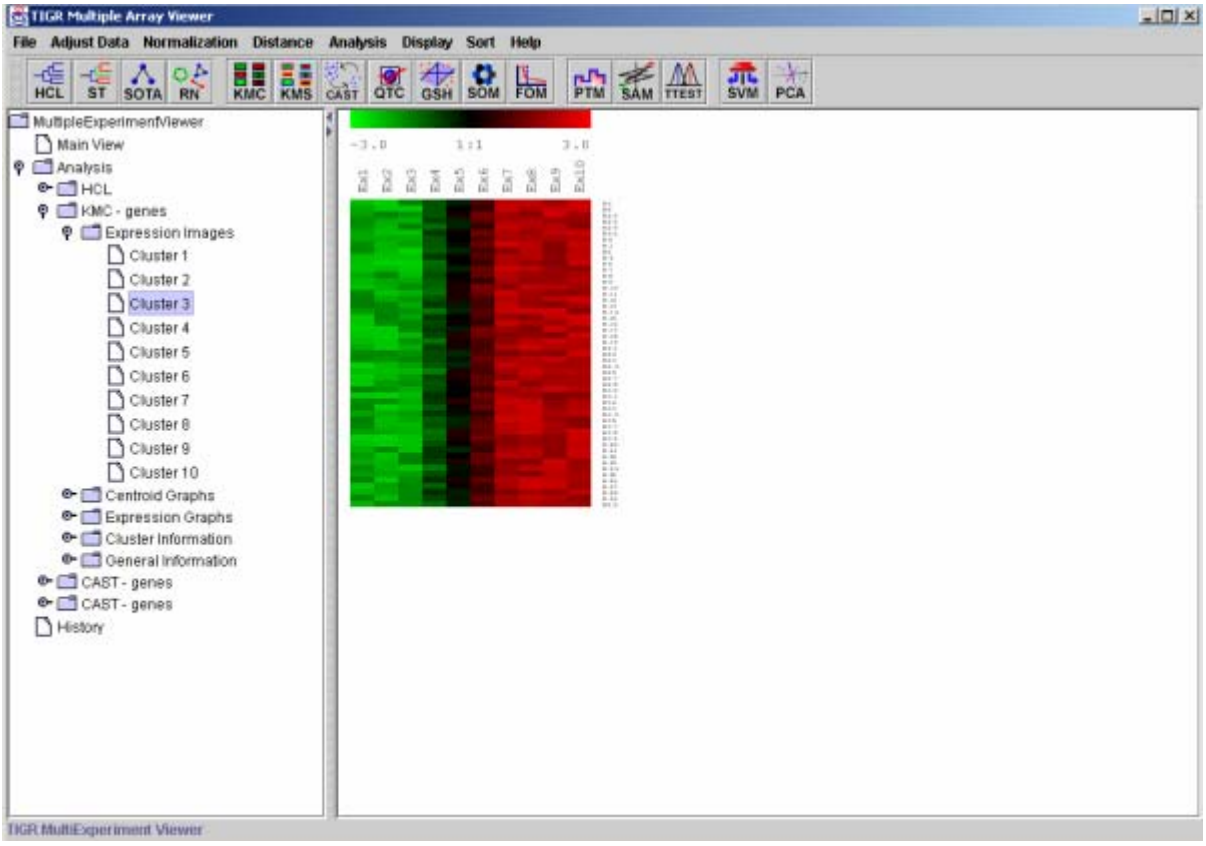


Da qui è possibile eseguire qualsiasi tipo di clustering e visualizzare i risultati in differenti modi a seconda del tipo di algoritmo utilizzato.

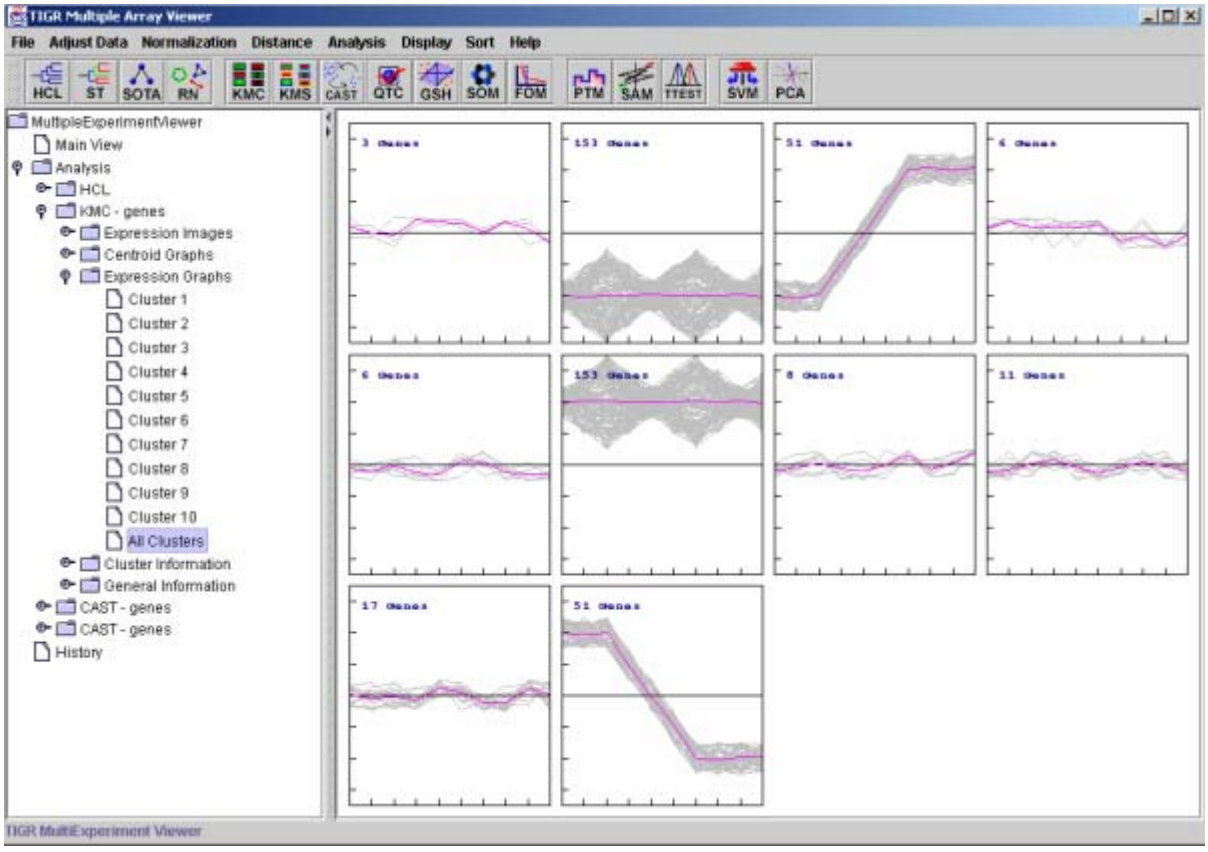
Il HCL produce esclusivamente un'immagine associata ai dendrogrammi, dalla quale possiamo scegliere e salvare i diversi cluster cliccando su di un nodo.



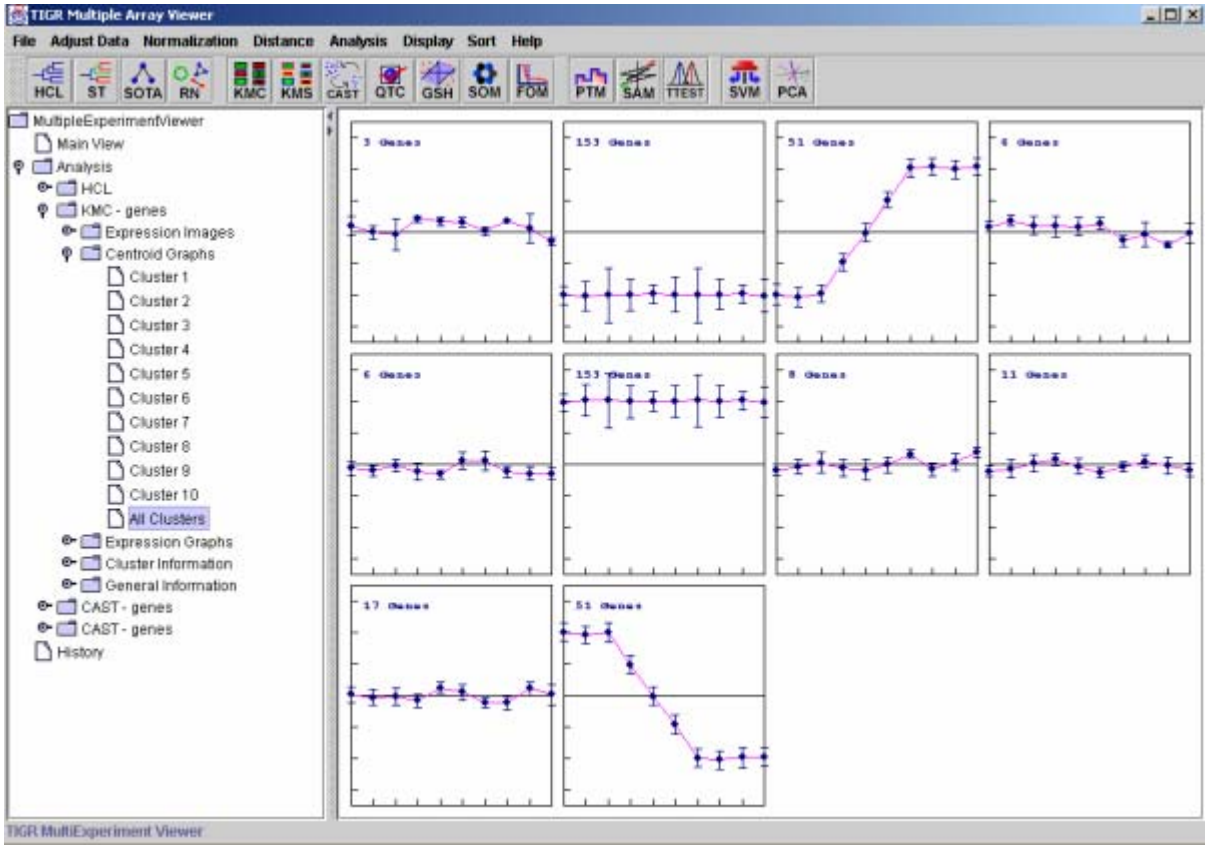
Il KMC produce un’immagine per ogni cluster



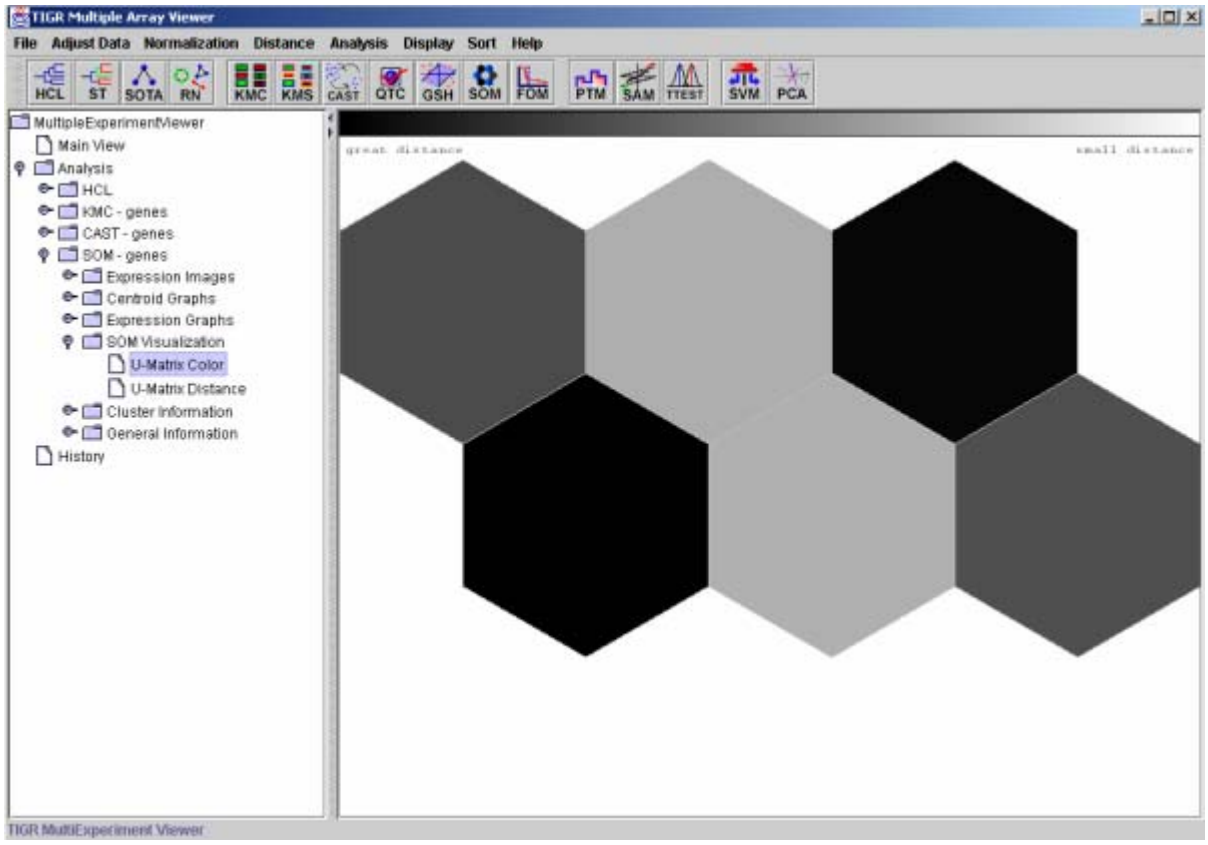
Ed inoltre anche le differenti rappresentazioni grafiche Expression Graphs



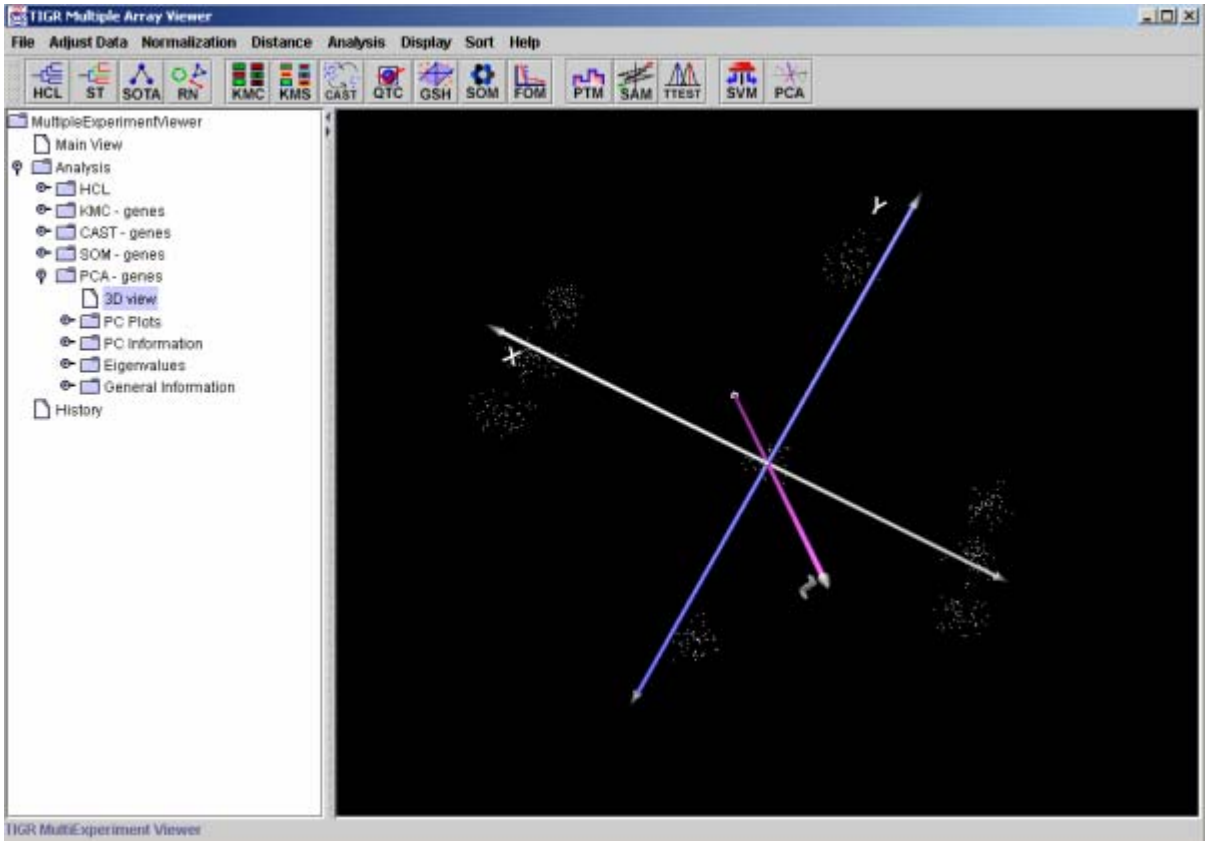
e Centroid Graphs



Tutto ciò è prodotto anche dal SOM che in più produce anche il tipo di visualizzazione della griglia generata in cui ogni cluster è disegnato di un colore che rappresenta la sua vicinanza agli altri.



E dal PCA che inoltre produce la visualizzazione 3D delle prime 3 principal component



E’ possibile salvare tutti i cluster e caricarli successivamente, cambiare i colori delle visualizzazioni e dei grafici, aggiustare e normalizzare i dati ed eseguire t-test e permutazioni.
Bisogna solo scoprire tutte le funzionalità e le novità di questo performantissimo software.
Buon Divertimento!