# UNIVERSITA' DEGLI STUDI DI NAPOLI
# "FEDERICO II"
# FACOLTA' DI INGEGNERIA
## CORSO DI LAUREA IN INGEGNERIA INFORMATICA


TESI DI LAUREA

# ANALISI DI SOFTWARE FREEWARE PER MICROARRAY


**Relatore**
**Ch.mo Prof. Ing.**
**Antonio D'Acierno**

**Candidato**
**Domenico Scialdone**
**Matr. 041/000892**


**ANNO ACCADEMICO 2002-2003**

# Indice

Indice 1
CAPITOLO I 4
Breve introduzione ad elementi di biologia, cellule, molecole, geni, bioinformatica e functional genomics. 4
1.1 Organismi e cellule 4
1.2 Le molecole delle cellule 6
1.3 Cromosomi, geni e genomi 14
1.4 Bioinformatica e functional genomics 22
CAPITOLO II 26
La tecnologia Microarray 26
2.1 Introduzione 26
2.2 L'esperimento microarray 27
2.3 L'analisi dei dati : Image quantify 29
2.4 L'analisi dei dati : algoritmi di normalizzazione e di data mining 31
2.5 L'analisi dei dati: la visualizzazione delle reti di interazione tra i geni 39
CAPITOLO III 41
Software per l' Image Quantify 41
3.1 Introduzione 41
3.2 Dapple: Image Analysis Software for DNA Microarrays. 41
3.3 F-scan 44
3.4 P-scan (Peak quantification using Statistical Comparative Analysis) 47
3.5 GridGrinder 49
3.6 ScanAlyze 2 53
3.7 TIGR Spotfinder 55
3.8 Spot 2.0 57
CAPITOLO IV 59

Software per la normalizzazione ed il data mining 59
4.1 Introduzione 59
4.2 Cluster and TreeView 60
4.3 Genesis 62
4.4 J-express 64
4.5 MAExplorer - MicroArray Explorer 66
4.6 TIGR Multiple Experiment Viewer (MEV) 68
4.7 AMIADA (Analysis of Microarray Data) 70
4.8 R-maanova 72
4.9 Genecluster 73
4.10 Clustfavor 75
CAPITOLO V 77
Software per la visualizzazione delle reti di interazione tra geni 77
5.1 Introduzione 77
5.2 Cytoscape 77
5.3 GenMAPP (Gene MicroArray Pathway Profiler) 80
5.4 Osprey Network Visualization System 82
CAPITOLO VI 84
Software web-based 84
6.1 Introduzione 84
6.2 Engene 84
6.3 Expression profiler 87
6.4 Gene Expression Data Analysis Tool (GEDA) 89
6.5 GEPAS (Gene Expression Pattern Analysis Suite) 91
6.6 Dynamic signaling map 93
CAPITOLO VII 95
Conclusioni 95
Bibliografia 98

# CAPITOLO I

# Breve introduzione ad elementi di biologia, cellule, molecole, geni, bioinformatica e functional genomics.

## 1.1 Organismi e cellule

Tutti gli organismi viventi sono composti da cellule, queste di solito sono troppo piccole per poter essere viste ad occhio nudo ma sono grandi abbastanza per un microscopio ottico. Ogni cellula è un sistema complesso formato da molti elementi ed è racchiusa da una membrana. Gli organismi possono essere unicellulari e multicellulari; i batteri ed il lievito di birra sono esempi di organismi unicellulari, mentre un cane un gatto o un essere umano sono esempi di organismi multicellulari.

In un corpo umano si stima che siano presenti circa $6^{13}$ cellule, di circa 320 tipi differenti. Il numero di tipi differenti non è ben definito perché questo dipende dal livello di dettaglio che vogliamo considerare per la classificazione; al limite se consideriamo il numero di molecole presenti in una cellula nessuna sarà uguale all'altra.

La grandezza delle cellule in un corpo umano può variare dal tipo di cellula e dalle circostanze di osservazione, un globulo rosso ha un diametro di circa 5 micron (0,005 mm) mentre alcuni neuroni sono lunghi circa un metro ( dalla spina dorsale alla gamba). In generale per animali e piante il diametro tipico di una cellula è compreso tra i 10 ed i 100 micron.

In natura ci sono due tipi differenti di organismi e di cellule rispettivamente: gli eucarioti (dal greco *eu* "buono" e *kàrion* "nucleo") ed i procarioti (dal greco *pro* "prima di" e *kàrion* "nucleo"). La distinzione tra procarioti ed eucarioti è piuttosto importante perché molti componenti costitutivi e processi vitali nei due tipi di cellule sono molto diversi. Si ritiene che queste differenze siano il risultato di differenti percorsi evolutivi. Esempio di procarioti sono i batteri

mentre il lievito di birra, i funghi, l'erba, gli alberi, i cani, i gatti e gli esseri umani sono esempi di eucarioti.

Un altro tipo di corpuscolo molto importante in biologia ed in genetica presente in natura è il *virus*. I virus non sono propriamente degli organismi viventi ma quando sono all'interno di una cellula di un organismo vivente mostrano alcune caratteristiche degli organismi viventi. I virus non possono essere osservati con un microscopio ottico, sono troppo piccoli (le loro dimensioni variano da 0,05 e 0,1 micron), ma con un microscopio elettronico si riesce ad osservare la loro struttura.

Le cellule procariotiche hanno una struttura più semplice (ad esempio non hanno il nucleo) e sono più piccole (la grandezza tipica è di 1 micron) delle cellule eucariotiche, ma questo non significa che siano meno importanti. Molti batteri che si trovano nella bocca e nell'apparato digerente degli esseri umani sono necessari per condurre una vita normale.

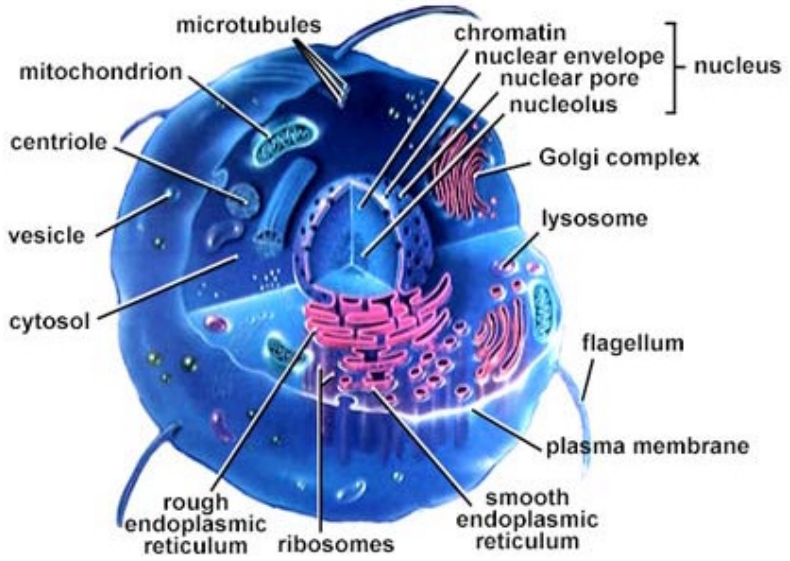



*Figura 1: modello di una cellula eucariotica (figura tratta da: On-line Biology Book).*

Le cellule eucariotiche hanno un nucleo, che è separato dal resto della cellula da una membrana, che contiene i cromosomi, che sono i portatori del materiale genetico. Ci sono inoltre altri elementi nella cellula racchiusi da membrane detti "organelli" che sono specializzati in svariati compiti. I mitocondri in particolare

sono specializzati nella produzione di energia (respirazione cellulare); c'è una teoria che dice che i mitocondri sono dei procarioti che vivono negli eucarioti. Nelle piante inoltre sono presenti i cloroplasti che producono zucchero utilizzando la luce. L'area della cellula esterna al nucleo e agli organelli è detta citoplasma. La membrana cellulare infine è una struttura complessa che ha una effettiva funzione di separazione della cellula dall'ambiente esterno, regola il flusso di cibo, energia ed informazioni da e verso la cellula.

Una caratteristica essenziale nella maggior parte delle cellule viventi è la loro abilità a crescere in un ambiente appropriato e ad effettuare la divisione cellulare; la crescita di una singola cellula e la sua susseguente divisione è chiamato *ciclo cellulare*. I procarioti ed in particolare i batteri sono estremamente efficienti nel moltiplicarsi; è probabile che la selezione naturale di questi organismi unicellulari abbia favorito quelli che erano più abili a crescere e a moltiplicarsi. Gli organismi multicellulari invece iniziano tipicamente la loro vita come cellula singola, risultato della fusione di una cellula sessuale maschile e di una femminile. Questa cellula iniziale cresce, si divide e si diversifica in vari tipi di cellula, per formare i tessuti e, negli eucarioti superiori, gli organi. Il processo di divisione e differenziazione è importante e delicato ed è controllato dalla cellula stessa. Malfunzionamenti in questo processo portano alla crescita incontrollata di cellule difettose e quindi ai tumori.

## 1.2 Le molecole delle cellule

Le cellule sono formate da molecole che sono state suddivise in quattro classi: le molecole piccole, le proteine, il DNA e l'RNA. Queste ultime tre sono conosciute anche con il nome di macromolecole biologiche.

**Le molecole piccole.**

Queste possono sia essere i blocchi costituenti delle macromolecole sia avere un ruolo indipendente, come essere un trasmettitore di segnali o una fonte di energia o materiale per le cellule. Qualche esempio importante oltre all'acqua sono lo zucchero, gli acidi grassi, gli amminoacidi e i nucleotidi. Ad esempio le membrane presenti in una cellula sono costruite a partire dagli acidi grassi, in cui

le macromolecole sono incastrate. Un altro esempio sono gli amminoacidi che sono i blocchi costitutivi delle proteine; ci sono 20 differenti amminoacidi ( ad essere più precisi ci sono 19 amminoacidi e uno che ha una struttura lievemente differente).

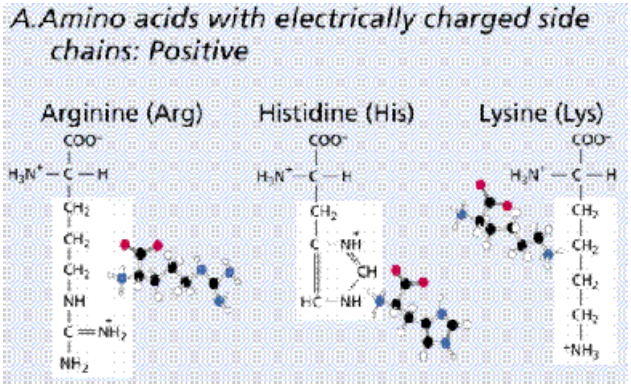



*Figura 2: struttura di alcuni amminoacidi.*

Questi sono tre esempi di amminoacidi. Essi hanno una parte comune e una parte differenziata che determina le loro proprietà. C'è una convenzione per cui ogni amminoacido è denotato da una lettera dell'alfabeto latino, ad esempio l'arginina dalla lettera R la istidina dalla lettera H la lisina dalla L e cosi via.

**Le proteine.**

Le proteine sono i principali blocchi costitutivi e funzionali della cellula, forniscono il 20% del peso di una cellula eucariotiche, il contributo più alto dopo l'acqua (70%). Ci sono molti tipi di proteine, le proteine strutturali, che possono essere pensate come gli elementi costitutivi base degli organismi, un esempio è il collagene, che è la principale proteina strutturale del tessuto connettivo e delle ossa; gli enzimi che forniscono (catalizzano) una moltitudine di reazioni biochimiche, come l'alterazione, l'unione o la divisione delle altre molecole, l'insieme delle reazioni e dei percorsi che fanno gli enzimi è chiamato metabolismo; le proteine della membrana cellulare che sono la chiave per la manutenzione dell'ambiente cellulare, regolando il volume delle cellule, l'estrazione o la concentrazione di piccole molecole dall'ambiente extracellulare e la generazione di gradienti ionici essenziali per il funzionamento delle cellule dei muscoli e dei nervi; un esempio è la pompa sodio/potassio.

Le proteine sono lunghe catene di amminoacidi ma a causa delle attrazioni e repulsioni tra gli atomi di questi amminoacidi hanno una complessa struttura tridimensionale (vedi le figure seguenti), e per un'analisi più accurata questa struttura è stata suddivisa in quattro livelli. Il primo livello è detto *struttura primaria*, è una struttura lineare in cui le proteine sono rappresentate dalla successione dei 20 simboli degli amminoacidi che le compongono, tale struttura è detta talvolta catena polipeptidica; la lunghezza delle proteine può variare da pochi a molte migliaia di amminoacidi componenti. Le *strutture secondarie* dipendono dalla sequenza degli amminoacidi e sono le eliche alfa , i filamenti beta e strutture meno regolari dette loops. Anche le strutture secondarie vengono in contatto tra loro e anche tra queste si stabilisce un equilibrio di forze intermolecolari che ne determina una struttura compatta e relativamente stabile , questa struttura e detta *struttura terziaria*. Una proteina può essere formata da più di una catena di amminoacidi, in questo caso si dice che abbiamo una *struttura quaternaria*. Per esempio l'emoglobina, è formata da quattro catene ognuna delle quali è in grado di legarsi con una molecola di ferro.

Le proteine sono troppo piccole per poter essere viste in un microscopio ottico. La grandezza tipica di una proteina è compresa tra i 3 e i 10 nanometri (1 nm= $10^{-9}$ m), scoprire la loro struttura è difficile e costoso.

Le immagini seguenti mostrano la struttura di una proteina usando particolari programmi di visualizzazione 3D, le immagini sono solo un modello della proteina non 'fotografie' che per le dimensioni, movimento e il principio di indeterminazione di Heisenberg non sono possibili.

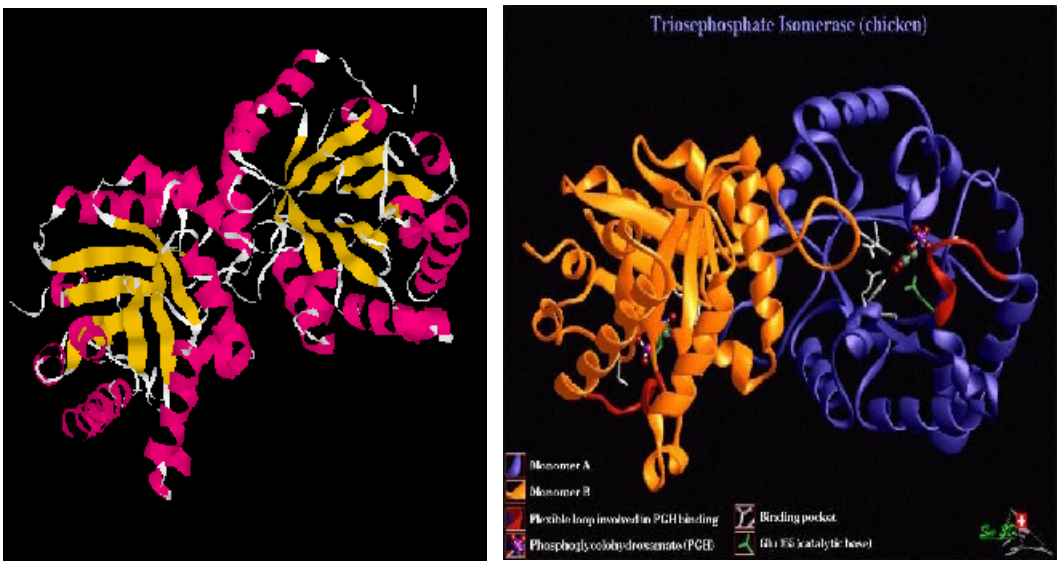



*Figura 3: modello di proteina (sinistra), il colore magenta è per le eliche alfa, il colore giallo per i filamenti beta; (destra) altro esempio di proteina in cui i colori evidenziano monomeri diversi*

Dall'analisi delle proteine si sono individuate circa 15000 strutture proteiche diverse, ma molte di queste sono molto simili tra loro e i biologi che si occupano di questo ritengono che siano circa 1500 le strutture differenti rappresentative scoperte finora, la somiglianza tra proteine è ritenuta anche un effetto dell'evoluzione.

Tutti e quattro i livelli strutturali delle proteine sono essenzialmente determinati dal primo livello, cioè dalla sequenza di amminoacidi, e dall'ambiente fisico-chimico in cui la molecola è posta. Predire la struttura della proteina a partire dalla sequenza degli amminoacidi è una delle sfide più importanti per la bioinformatica.

Per avere un'idea delle dimensioni relative tra cellule e proteine , le proteine sono 10000 volte più piccole delle cellule e in ogni cellula ci sono circa 2 miliardi di proteine.

Sebbene le forze come il legame idrogeno siano deboli individualmente, quando due o più macromolecole biologiche con forma complementare sono vicine, la somma di tutti questi piccoli legami possono far si che le molecole interagiscano fortemente. Tutte queste forze intermolecolari ed interazioni giocano un ruolo fondamentale nei processi vitali e sono alla base dei processi biologici. Per esempio molte proteine possono unirsi per formare grandi complessi proteici,

come l’ RNA polimerase II che nel lievito legge e trascrive le informazioni genetiche ed ha 10 sotto-unità.

**Il DNA.**

Il DNA è la principale molecola portatrice di informazioni in una cellula. Il DNA può essere a filamento singolo o doppio. Una molecola a singolo filamento è anche chiamata polinucleotide ed è una catena di molecole più piccole chiamate nucleotidi. Ci sono quattro tipi differenti di nucleotidi in due gruppi distinti. Le *purine,* adenina e guanina, e le *pirimidine,* citosina e timina. Queste sono chiamate di solito basi (in fatti le basi sono gli unici elementi distinguibili tra differenti nucleotidi) e denotati dalle loro lettere iniziali A,C,G e T (da non confondersi con gli amminoacidi).

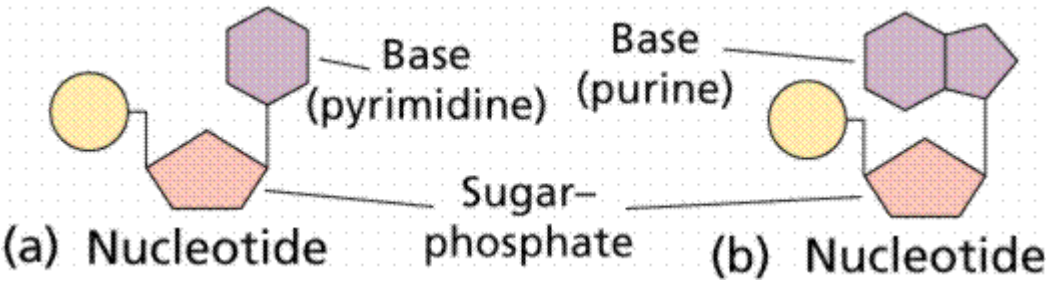



*Figura 4: nucleotidi (figura tratta da “On-Line Biology Book”)*

I nucleotidi possono essere collegati in qualsiasi ordine per formare un filamento, ad esempio come questo :

```
A-G-T-C-C-A-A-G-C-T-T
```

I filamenti possono essere di lunghezza qualsiasi e dato che i due estremi dei nucleotidi sono chimicamente differenti, le sequenze hanno una direzionalità, come nel caso seguente:

```
A->G->T->C->C->A->A->G->C->T->T
```

Due stringhe sono dette complementari se l'una è ottenuta dall'altra scambiando A con T e C con G e cambiando la direzione delle molecole in quella opposta, ad esempio:

```
T<-C<-A<-G<-G<-T<-T<-C<-G<-A<-A
```

è la stringa complementare di quella precedente.
Particolari coppie di nucleotidi possono formare deboli legami tra loro. A lega con T e C con G (per essere più precisi, due legami idrogeno possono essere formati da una coppia A-T, e tre legami idrogeno dalla coppia C-G). Sebbene queste interazioni siano deboli individualmente, quando due filamenti complementari si incontrano questi tendono ad unirsi come nel caso illustrato:

```
5′  C-G-A-T-T-G-C-A-A-C-G-A-T-G-C  3′
    | | | | | | | | | | | | | |
3′  G-C-T-A-A-C-G-T-T-G-C-T-A-C-G  5′
```

Gli estremi dei filamenti sono marcati con 5' e 3' per mettere in evidenza la loro direzionalità; per convenzione il DNA è di solito scritto con 5' alla sinistra e 3' alla destra con la striscia di codice in alto.
Le linee verticali tra le due stringhe rappresentano i legami tra nucleotidi anche se per la precisione si sarebbero dovute inserire due linee tra la coppia A-T e tre tra la coppia C-G come mostrato nelle figure seguenti. Le coppie A-T e C-G sono chiamate coppie base (pb). La lunghezza del DNA è misurata in numero di coppie base o nucleotidi che in questo contesto è la stessa cosa.

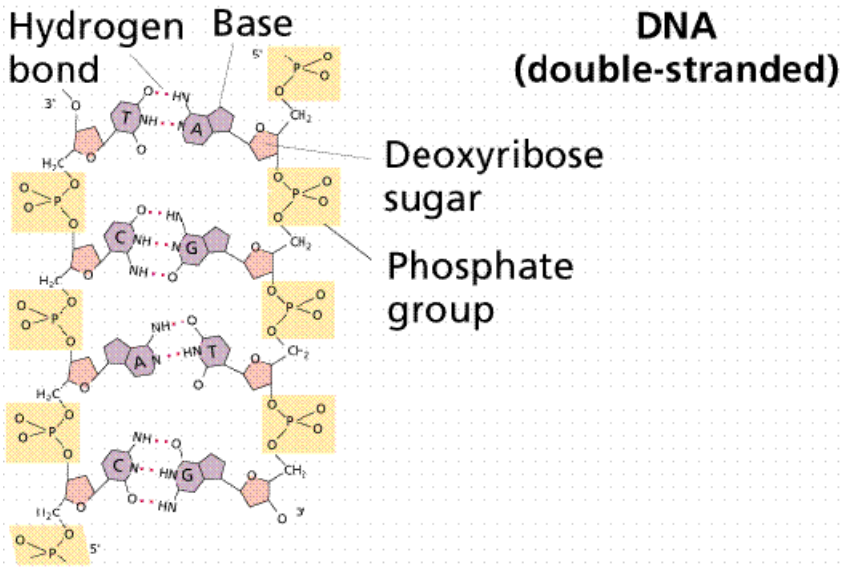



*Figura 5: struttura chimica del DNA (figura tratta da "On-Line Biology Book")*

Due filamenti complementari formano una struttura stabile, che assomiglia ad un'elica ed è conosciuta come la doppia elica del DNA. Sono necessarie circa 10 coppie base per un intero giro che è lungo circa 3.4 nm.

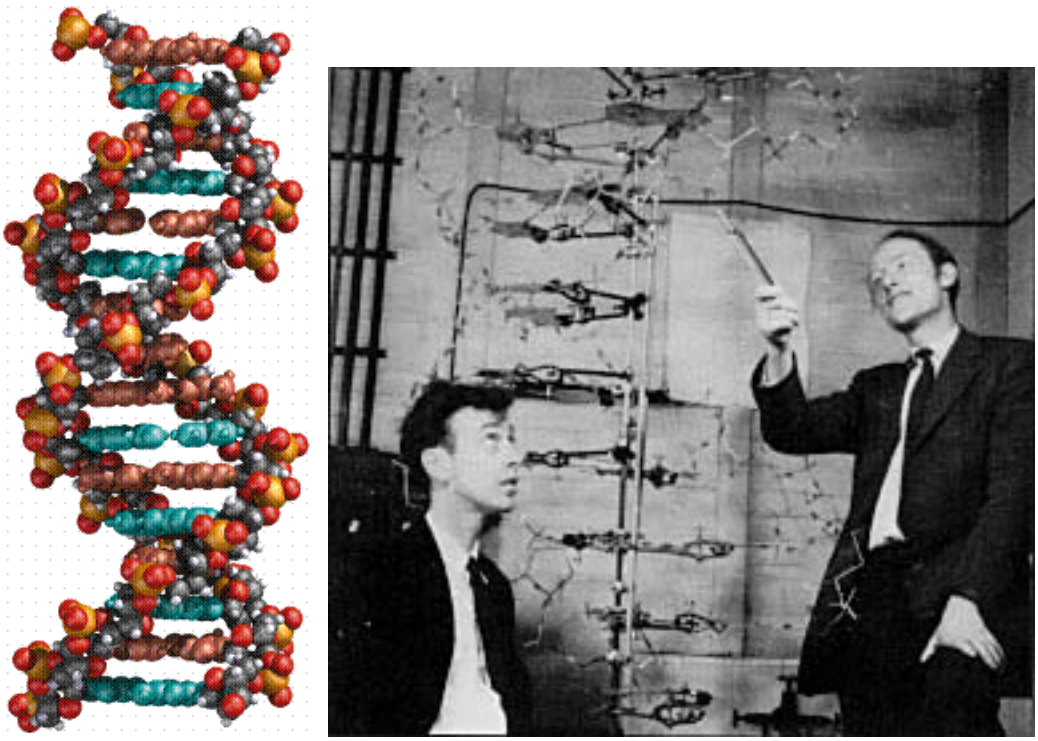


*Figura 6: Struttura tridimensionale a doppia elica del DNA (sinistra). James Watson e Francis Crik ed il loro modello di DNA (destra). (figure tratte da "On-Line Biology Book")*

Questa struttura fu mostrata per la prima volta a Cambridge nel 1953 da James Watson e Francis Crik (con l'aiuto di altri), per questa scoperta hanno avuto il premio Nobel.

Due filamenti complementari di DNA formano una struttura stabile quasi indipendentemente dalla sequenza di nucleotidi che li formano. Questo permette alle molecole di DNA di essere un mezzo perfetto per l'immagazzinamento delle informazioni. Ognuno dei filamenti determina l'altro quindi ai fini delle informazioni contenute è sufficiente fornire solo una sequenza di nucleotidi come nella stringa seguente CGATTCAACGATGC. La quantità massima di informazione che può essere codificata in questo tipo di molecole è di 2 bit per la lunghezza della sequenza. Dato che la lunghezza tra due nucleotidi è di 0.34 nm, la densità lineare di immagazzinamento delle informazioni è di $6^8$ bits/cm che è pari a circa 75 GB/cm.

La complementarietà dei due filamenti è usata per copiare le molecole di DNA in un processo conosciuto come 'replicazione del DNA' in cui una doppia elica di DNA è replicata in una coppia di molecole uguali: la doppia elica del DNA durante il processo si biforca e un nuovo filamento complementare è sintetizzato su ognuna delle parti per avere alla fine del processo due eliche identiche all'originale. In una cellula questo accade durante la divisione cellulare e una copia identica all'originale viene distribuita in ogni cellula.

Se la somma totale delle forze dei legami tra i nucleotidi complementari sono abbastanza intense da conservare la struttura, sono possibili anche nucleotidi spaiati tra i filamenti. Quindi molecole come questa

```
C-G-A-T-T-G-C-C-A-C-G-A-T-G-C
| | | ~ | | | ~ | | | ~ | | |
G-C-T-T-A-C-G-T-T-G-C-A-A-C-G
```

sono possibili sebbene abbastanza rare in una cellula vivente.

Se non ci sono abbastanza legami le forze tra i due filamenti possono diventare deboli e i due filamenti dividersi. Il numero di legami necessari a mantenere i

due filamenti legati dipende dalla temperatura e da altri fattori ambientali. Il DNA che non è più in forma elicoidale viene detto denaturato.

**L‘ RNA.**

L’RNA come il DNA è costituito da nucleotidi. Ma invece della timina T ha in alternativa l’ *uracile* U, che non si trova nel DNA. A causa di questa differenza l’ RNA non forma una doppia elica, ma è in forma di filamento singolo anche se può avere una struttura spaziale molto complessa dovuta ai legami tra nucleotidi complementari dello stesso filamento. L’RNA ha varie funzioni nella cellula, alcune di queste discusse nella sezione successiva, ad esempio, l’mRNA e il tRNA sono due tipi di RNA funzionalmente differenti ma che servono entrambi per la sintesi di proteine.

L’RNA può legarsi in modo complementare ad una singola stringa di DNA dove T è sostituito da U, così si possono formare molecole come questa

```
C-G-A-T-T-G-C-A-A-C-G-A-T-G-C   DNA
| | | | | | | | | | | | | |
G-C-U-A-A-C-G-U-U-G-C-U-A-C-G   RNA
```

che giocano un ruolo importante nei processi vitali e nelle biotecnologie. C’è un’ipotesi che dice che le prime forme di vita sulla terra si sono basate solo sull’RNA.

L’RNA può codificare le informazioni genetiche, è replicabile, forma complesse strutture tridimensionali e può anche agire da catalizzatore per alcune reazioni chimiche legate alla riproduzione.

## 1.3 Cromosomi, geni e genomi

In una cellula ci sono una o più molecole di DNA ad elica organizzate in cromosomi. Negli Eucarioti i cromosomi hanno una struttura complessa e il DNA gira intorno a proteine strutturali chiamate istoni. Un essere umano ha 23 coppie di cromosomi, che sono grandi abbastanza da poter essere viste con un microscopio ottico. La lunghezza totale del DNA in una cellula umana, se potessimo svolgerla sarebbe più di un metro.

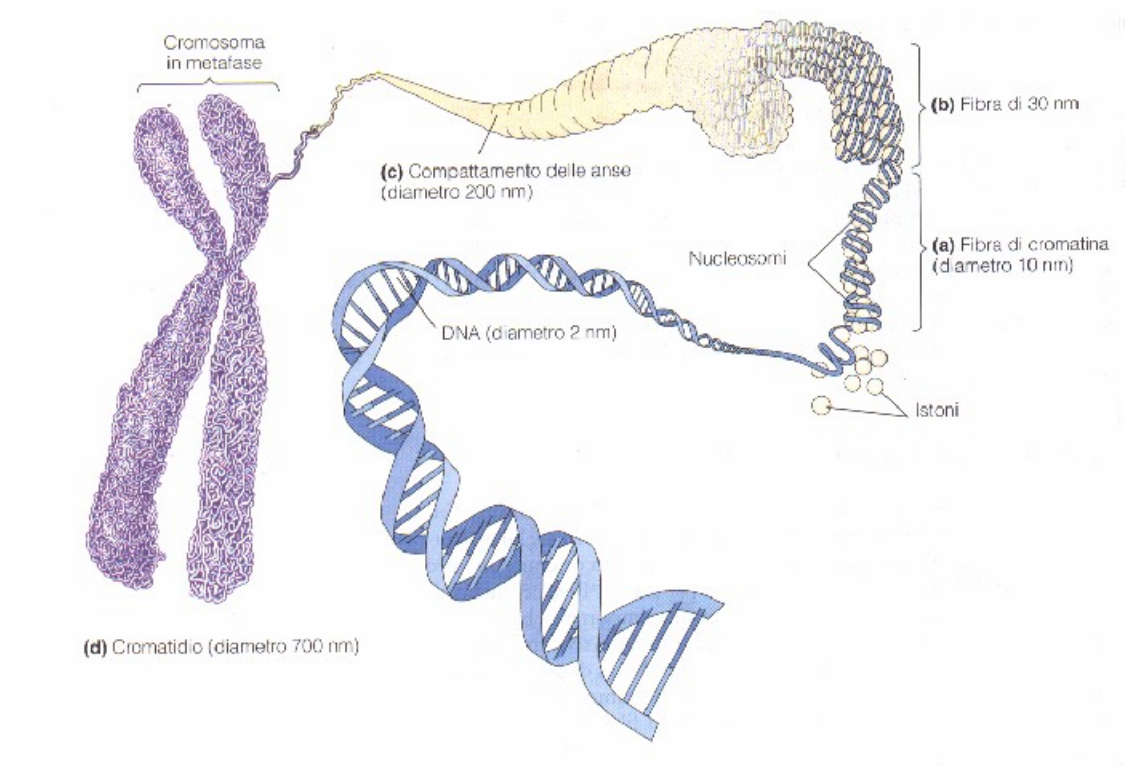



*Figura 7: tappe del compattamento dei cromosomi eucariotici (figura tratta da "Genetica,analisi e principi")*

Anche i mitocondri contengono DNA, ma questa quantità è minuscola rispetto a quella dei cromosomi. Il DNA dei cromosomi e dei mitocondri forma il genoma dell'organismo. Tutti gli organismi hanno il genoma e si ritiene che quasi tutte le informazioni ereditarie dell'organismo siano codificate in esso. Negli eucarioti i cromosomi sono nel nucleo (tranne i genomi dei mitocondri), racchiusi dalla membrana nucleare, mentre nei procarioti il DNA è distribuito nella cellula che è priva di nucleo. Tutte le cellule di un organismo contengono il medesimo patrimonio genetico, tranne qualche eccezione, come le cellule adibite alla riproduzione.

C'è un meccanismo molecolare nelle cellule che mantiene intatto il DNA in entrambe le stringhe: se una stringa è danneggiata, è riparata usando l'altra come riferimento. E' importante che un danno al DNA, causato ad esempio da cause esterne come le radiazioni, come la rottura di una o entrambi i filamenti o il disaccoppiamento delle basi porti soprattutto all'interruzione del meccanismo di replicazione del DNA. Se il danno al DNA non è riparato si hanno due

alternative, la morte della cellula o il tumore. I cambiamenti nel DNA sono dette mutazioni.

La dimensione del genoma differisce abbastanza considerevolmente da organismo ad organismo come si può vedere dalla tabella seguente:

| organismo | Numero di cromosomi | Dimensioni del genoma in coppie base (bp) |
|---|---|---|
| batterio | 1 | ~400,000 - ~10,000,000 |
| lievito | 12 | 14,000,000 |
| verme | 6 | 100,000,000 |
| mosca | 4 | 300,000,000 |
| erba | 5 | 125,000,000 |
| uomo | 46 | 3,000,000,000 |

Queste dimensioni  determinano il limite superiore delle informazioni genetiche di un organismo.

La determinazione della sequenza delle quattro lettere per un dato DNA è detto sequenziamento del DNA. Il primo genoma intero di un batterio fu individuato nel 1995. Il genoma del lievito nel 1997, il verme nel 1999, la mosca nel 2000, l'erba nel 2001, e nel 2001 più del 90% del genoma umano è stato individuato.

L'individuazione dei genomi di batteri è diventata routine ed è fatta per la maggior parte da robot e completata da ricercatori umani. Il problema principale è la minimizzazione dei costi per lettera e la massimizzazione della velocità mantenendo la qualità. L'individuazione di genomi di dimensioni rilevanti come quello umano è ancora difficile sebbene i maggiori problemi sono computazionali.

I robot sono in grado di individuare solo piccole sequenze alla volta che devono poi essere assemblate con l'utilizzo di algoritmi di assemblaggio. La maggiore difficoltà è che il genoma degli eucarioti superiori come l'uomo ha molte sottosequenze ripetute che rendono l'assemblaggio complicato e questo comporta un notevole intervento umano nel progetto di sequenziamento. I genomi contengono geni, la maggior parte dei quali codifica proteine.

**I geni e la sintesi delle proteine.**

Ci sono molte discussioni tra i biologi per trovare una definizione di gene che sia esauriente per tutti. Per i nostri scopi possiamo introdurre questa: "Un gene è una stringa continua di DNA, dalla quale un complesso meccanismo molecolare può leggere informazioni (codificate come una stringa di A, T, G e C) e formare un particolare tipo di proteina oppure poche proteine differenti".

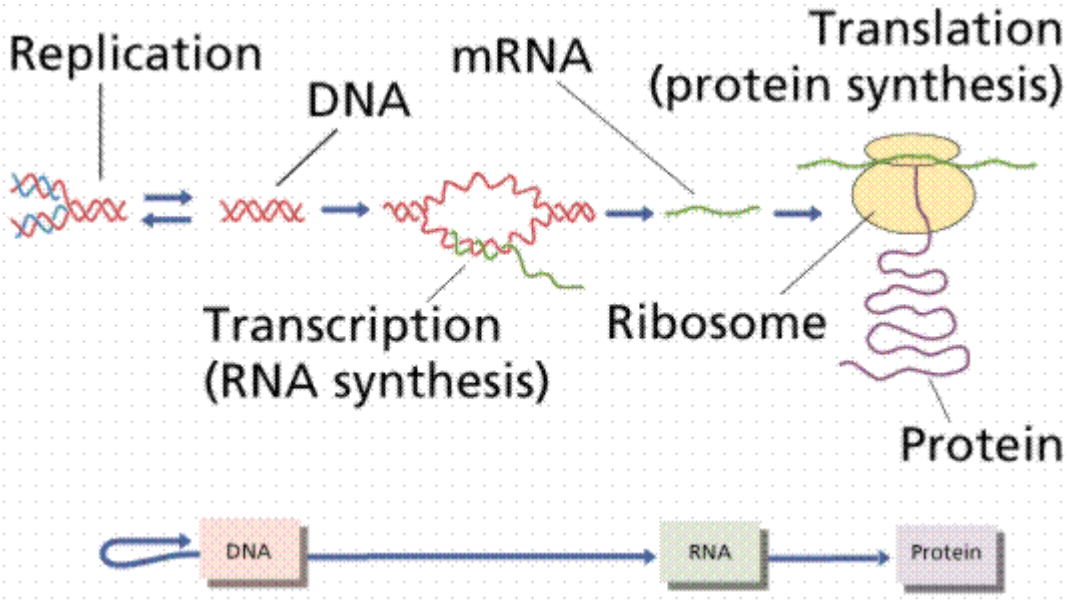



*Figura 8: Processo di sintesi delle proteine.*

Questa definizione non è precisa e per capirla meglio abbiamo bisogno di descrivere il meccanismo molecolare di creazione delle proteine basato sulle informazioni codificate nei geni. Questo processo è chiamato sintesi proteica ed ha tre fasi essenziali: trascrizione, splicing e traduzione.

Nella fase di trascrizione un filamento di DNA è copiato in un pre-mRNA ( pre per preliminare e m per messaggero ) dal complesso proteico RNA polimerase .

Nel processo i due filamenti di DNA legati in doppia elica vengono slegati e l'informazione è letta solo da un filamento .

Nella fase di splicing ("taglio e saldatura") alcuni tratti di pre-mRNA, chiamati introni, vengono rimossi e le sezioni rimanenti chiamate esoni vengono unite. La rimozione degli introni è una conseguenza del modo di come i genomi degli eucarioti sono organizzati. Il DNA genomico che corrisponde al codice del gene non è continuo, ma consiste di introni ed esoni. Gli esoni sono la parte del gene che codifica le proteine e sono interspaziati da introni che devono essere rimossi nella fase di splicing. Il numero e le dimensioni degli introni ed esoni differisce considerevolmente tra un gene ed un altro e anche tra specie diverse. Solo alcuni geni nel lievito hanno gli introni, mentre negli uomini ci sono in media circa 4 introni per gene, e la lunghezza media di un esone è di 150 bp e circa 3400 bp per un introne. I geni dei procarioti non hanno introni e la fase di splicing non è presente. Il risultato della fase di splicing e l' mRMA. Molti geni degli eucarioti hanno differenti varianti della fase di splicing dove ad esempio alcuni pre-mRNA producono differenti mRNA, questo processo è detto splicing alternativo.

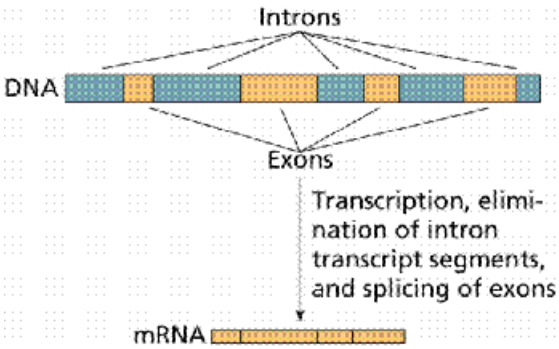



*Figura 9: fase di splicing (figura tratta da On-Line Biology Book)*

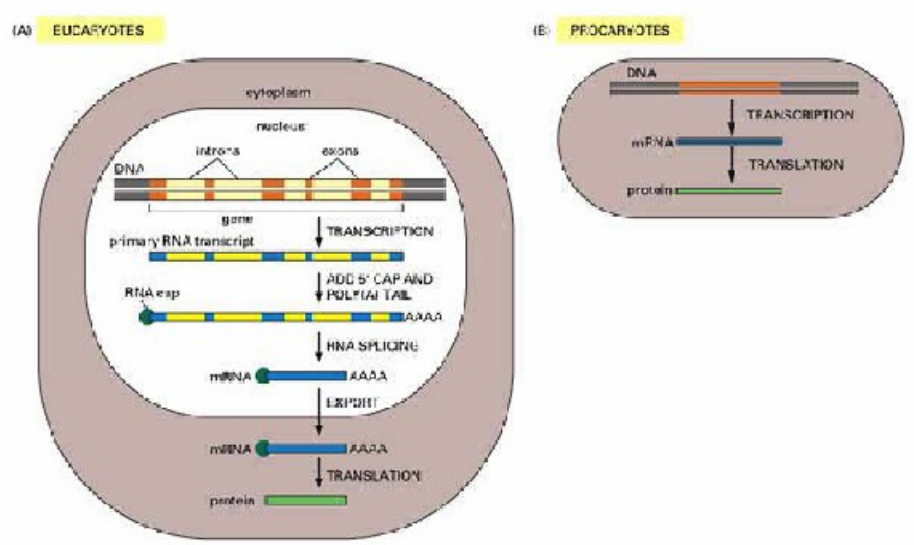



*Figura 10: fase di splicing negli eucarioti (a) e nei procarioti (b).*

La fase di traduzione è il processo di costruzione delle proteine attraverso l'unione di amminoacidi nell'ordine dato dal codice dell'mRNA. L'ordine degli amminoacidi è determinato da 3 nucleotidi adiacenti nel DNA. Ogni tripletta è chiamata codone ed è il codice di un solo amminoacido. Ci sono 64 codoni e solo 20 amminoacidi quindi il codice è ridondante, ad esempio la istidina è codificata da CAT e CAC.

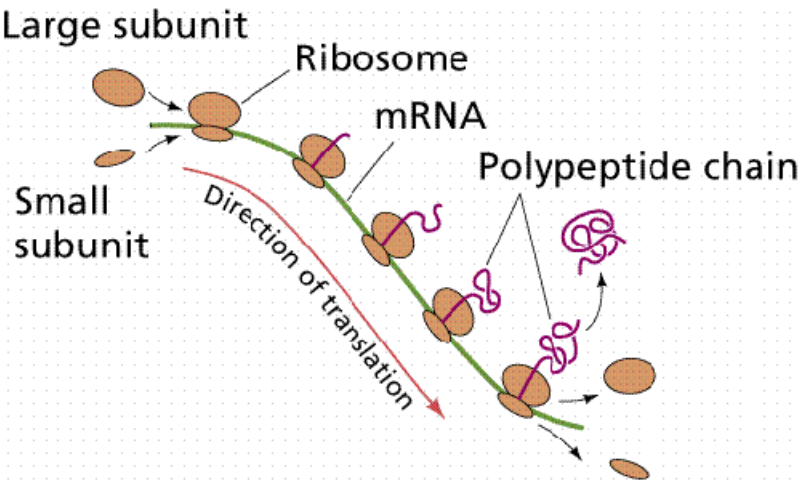



*Figura 11: fase di traduzione e formazione delle proteine.(figura tratta da On-Line Biology Book)*

Nel citoplasma l'mRNA forma un complesso con i ribosomi e viene formata materialmente la proteina. Ci sono differenti portatori o molecole di tRNA

ognuna trasportante uno specifico amminoacido al ribosoma e identificante uno specifico codone del mRNA, l'amminoacido trasportato dal tRNA e aggiunto alla nascente proteina. La fine della traduzione corrisponde alla parte finale del gene e il prodotto finale è la proteina, la sequenza della quale è in corrispondenza con la sequenza del codice di mRNA. Le proteine possono essere poi successivamente modificate ad esempio aggiungendo zucchero oppure spaccandosi in proteine più piccole.

Inizialmente si credeva nel paradigma una proteina, un gene, ma adesso con la scoperta dello splicing alternativo e la spaccatura delle proteine, un gene può produrre più proteine differenti. Ci sono anche geni che non codificano proteine ma RNA ( ad esempio il tRNA ).

**La predizione e conteggio dei geni.**

Una domanda interessante: data una sequenza di DNA genomico, possiamo dire dove si trova il gene nel cromosoma? La risposta è che possiamo, sebbene l'accuratezza di una tale predizione non è molto alta. La maggior parte della conoscenza di queste predizioni viene da geni identificati sperimentalmente. Ci sono circa 6700 geni sperimentalmente confermati nel genoma umano (per essere più accurati 6700 cDNA verificati). La predizione dei geni è un problema importante per la bioinformatica e ci sono vari algoritmi che tentano questa predizione usando i geni conosciuti come dati di partenza. Una tecnica algoritmica popolare usata nella predizione di geni sono i modelli nascosti di Markov (HMM). La predizione automatica di geni è stata importante nel dare senso ai dati non completi o non terminati dei vari progetti genoma: la sequenza totale è assemblata a partire da marcatori genetici conosciuti sui vari cromosomi.

Anche se conoscessimo dove i geni sono nel genoma, non è ovvio come contarli. A causa dell'esistenza geni sovrapposti e delle varianti, è difficile definire quale parte del DNA può essere legato allo stesso o a differenti geni. Tuttavia, anche se in modo approssimato , un conteggio dei geni in un organismo si è riusciti a farlo ed alcuni risultati hanno dato delle sorprese.

| Organismo | Numero di geni predetti | Parte del genoma che codificante (esoni) |
|---|---|---|
| E.Coli (bacterio) | 5000 | 90% |
| lievito | 6000 | 70% |
| verme | 18000 | 27% |
| mosca | 14000 | 20% |
| erba | 25500 | 20% |
| uomo | 30000 | < 5% |

Una delle sorprese è il numero relativamente piccolo di geni in un genoma umano rispetto a quello di un verme. Prima che la maggior parte del genoma umano venisse sequenziato si credeva che vi fossero circa 100000 geni in un uomo. Infatti alcuni esperimenti ancora dicono che ci devono essere al massimo 40000- 50000 geni nel genoma umano, e che 30000 riflette solo l'incapacità di predizione dei mezzi automatici. Ancora, sembra che non c'è una semplice correlazione tra l'intuitiva (e non ben definita) complessità di un organismo ed il numero dei geni nel suo genoma ( ad esempio una mosca che intuitivamente sembra più complessa di un verme ha meno geni di questo). La ragione del basso numero di geni nel genoma umano potrebbe essere l'elevato numero di varianti che i geni umani hanno nella sintesi proteica, ma questo non è stato ancora provato.

La presenza del 95% di DNA non codificante nel genoma umano rimane un mistero. Ci sono molte ipotesi che cercano di spiegare il fenomeno, ma nessuna è generalmente accettata.

**La somiglianza dei genomi, SNPs e il confronto di genomi.**

Di chi è il genoma del progetto di sequenziamento del DNA umano ? Molti campioni di DNA di persone anonime furono prelevati e tra questi se ne scelse uno, ma l'individuo particolare non è molto importante perché si ritiene che i genomi dei vari individui siano equivalenti al 99,9% ed in media solo un nucleotide su mille è differente nei genomi di due diversi individui. Quindi

possiamo parlare in termini di genoma umano in generale. Quello 0.1% di differenze è usato per il riconoscimento degli individui: si analizzano le variazioni in parti non codificanti del genoma e si producono schemi che possono distinguere in modo affidabile due individui differenti, si possono avere dei problemi però per la distinzione di due gemelli identici.

Variazioni particolarmente importanti nei genomi individuali sono i *singles nucleotide polymorphisms* o SNPs, che possono presentarsi sia nelle parti codificanti che in quelle non codificanti del genoma. Le SNPs sono variazioni di sequenze di DNA che si presentano quando una singola base (A,C,T,G) è alterata così che differenti individui possono avere lettere differenti in posizioni isotope. Particolari nucleotidi nelle posizioni SNP dentro o vicino ai geni possono influenzare la proteina prodotta dal gene. Alcune varianti proteiche ‘ anormali ‘ (varianti mutanti) sono causa di malattie genetiche. SNPs possono essere responsabili di molte differenze ereditarie tra gli individui e le variazioni di SNP possono indicare la predisposizione a malattie genetiche. Ad esempio è stato provato che certe combinazioni di SNPs sono presenti in individui con la malattia di Alzheimer.

I progetti di sequenziamento hanno rivelato che i genomi di organismi che sembrano molto differenti possono essere abbastanza simili. Si stima che la differenza tra il genoma di un uomo e quello di uno scimpanzé è solo dell’1-3%, mentre tra un uomo ed un topo solo dell’5-15% , in dipendenza di come la similitudine è definita e misurata. Queste similitudini derivano dalle similitudine dell’evoluzione di questi mammiferi. E’ possibile costruire un albero genetico dell’evoluzione delle proteine, dei geni e degli organismi basato sul confronto delle sequenze del DNA.

## 1.4 Bioinformatica e functional genomics

La necessita di archiviare l'enorme mole di dati che la moderna ricerca in biologia molecolare produce continuamente grazie al progresso tecnologico recente, spinge alla creazione, gestione e manutenzione di banche dati specializzate, dove a competenze di tipo strettamente informatico sono

affiancate competenze di tipo biologico per l'archiviazione dei dati in modo corretto, ragionato e di facile reperibilità per la comunità scientifica.

Questa prima applicazione della informatica alla biologia ha dato origine ad una nuova disciplina, la bioinformatica.

Questa disciplina è caratterizzata dal fatto che le informazioni di tipo biologico da trattare sono diventate quantitativamente molto rilevanti e non è più possibile analizzarle senza l'ausilio di sistemi informatici; si pensi al progetto genoma umano che ha permesso l'identificazione dei circa 30000 geni dell'uomo, solo con l'ausilio di un computer è possibile gestire efficientemente tutti questi dati.

Le applicazioni classiche della bioinformatica sono l'archiviazione, la ricerca, la predizione o l'analisi, della composizione o della struttura, delle molecole biologiche, come il DNA e le proteine.

Con il completamento del sequenziamento del genoma di molte specie sia animali che vegetali tra cui quello dell'uomo, le applicazioni della bioinformatica si sono ampliate e diversificate.

Una prima applicazione è quella di confrontare i genomi delle varie specie per osservarne differenze e similitudini per studiarne l'evoluzione, questo ramo della bioinformatica è detto *comparative genomics*.

Un'altra tecnologia sviluppata a partire dalla conoscenza dei genomi è quella dei *microarray* in cui viene misurato contemporaneamente per un gran numero di geni il livello di espressione del gene, per vari tessuti ed in molteplici condizioni; questa tecnologia è di supporto ad un'altro ramo della bioinformatica, quello della *functional genomics* in cui si cerca di capire le funzioni e le interazioni dei geni e dei loro prodotti.

La functional genomics quindi punta alla scoperta delle funzioni biologiche dei geni e a scoprire le interazioni di insiemi di geni e dei loro prodotti (proteine, RNA) nelle cellule sane e in quelle malate e come questo è collegato alle funzioni dell'organismo che si sta analizzando.

**Functional genomics: la quantità di proteine in una cellula.**

Le proteine in una cellula sono sintetizzate dai geni e il loro ciclo di vita può essere approssimativamente descritto come: sintesi, funzionalità e degradazione.

Nessuno realmente conosce quante proteine differenti sono sintetizzate dai circa 30000 geni in una cellula umana, ma causa dello splicing alternativo e alle modificazioni dopo la traduzione il numero delle proteine è apparentemente molto più alto del numero dei geni. Evidentemente non tutte le proteine devono essere presenti in una data cellula in un determinato momento.

La quantità di proteine dipende da diversi fattori: se il rispettivo gene è espresso oppure no, quanto intensamente (quanto velocemente) è espresso, se e quanto velocemente è tradotto e modificato, quanto è lunga la vita del mRNA e delle proteine. Studi sperimentali diretti sulla quantità delle proteine sono tecnicamente difficili, tuttavia grazie alla tecnologia microarray è possibile misurare l'abbondanza del mRNA (l'espressione del gene) per decine di migliaia di geni in parallelo in un singolo esperimento. La correlazione tra l'espressione del gene e la presenza delle rispettive proteine nella cellula non ancora chiara, ma in molti casi le stime delle proteine possono essere fatte a partire dall'espressione del gene.

Un'analisi interessante legata alle proteine in una cellula è di individuare le loro interazioni nel corso della vita della cellula.

**Functional genomics: la regolazione dei geni.**

Una questione importante ed interessante in biologia è come l'espressione del gene è accesa e spenta cioè come sono regolati i geni. Anche se quasi tutte le cellule in un organismo hanno un genoma identico, la differenziazione cellulare durante la vita dell'organismo sono dovute alle differenze nei contenuti delle espressioni dei geni e non nel genoma.

La regolazione dei geni negli eucarioti, non è ancora del tutto nota, ma ci sono prove che un ruolo importante è giocato da un tipo di proteine chiamato fattore di trascrizione. I fattori di trascrizione possono attaccarsi a parti specifiche del DNA, chiamate (combinazioni specifiche e relativamente brevi di A,T,C e G) promotori. Promotori specifici sono associati a geni particolari e sono generalmente non molto lontani dai rispettivi geni, tuttavia alcuni effetti regolatori possono essere localizzati anche 30000 basi lontano e questo rende difficile la definizione dei promotori.

I fattori di trascrizione controllano l'espressione dei geni collegandosi ai promotori e attivando così la trascrizione del gene o disattivandola. I fattori di trascrizione sono anche loro prodotti dai geni quindi a turno sono controllati anche essi da altri fattori di trascrizione. I fattori di trascrizione possono controllare molti geni e alcuni (probabilmente la maggior parte) sono controllati da una combinazione di fattori di trascrizione. Sono possibili cicli retroazionati, quindi possiamo parlare di reti di regolazioni dei geni. Capire, descrivere e modellare questa rete di regolazione dei geni è uno delle sfide più impegnative in functional genomics.

Ci sono più di 200 fattori di trascrizione conosciuti nel lievito, e più di 600 nei vermi e nelle mosche, e più di 1500 nell'erba. Ma il numero reale è probabilmente più alto, dato che più della metà dei geni previsti in questi organismi hanno un funzionamento sconosciuto, ci sono probabilmente dei fattori di trascrizione che devono essere ancora scoperti. Inoltre, i fattori di trascrizione non sono le uniche proteine a partecipare alla regolazione dei geni ed è risaputo che alcune regolazioni avvengono nella fase di traduzione. I microarray e i metodi computazionali stanno giocando un ruolo sempre maggiore nel cercare le reti di regolazioni con la tecnica del reverse engeneering a partire da molte osservazioni. Si deve tener presente comunque che in realtà la regolazione dei geni è un processo stocastico non deterministico.

La biologia molecolare tradizionale ha seguito fino ad ora un approccio detto riduzionista concentrando lo studio su di un solo o molto pochi geni per ogni progetto di ricerca. Da quando i genoma sono sequenziati la situazione sta cambiando in quello che viene detto approccio sistemico.

Ci possiamo iniziare a chiedere quanti geni sono espressi nei differenti tipi di cellule, quali geni sono espressi in tutti i tipi di cellule, quali sono le regole funzionali di questi geni, quanto grande è l'universo funzionale dei geni, quanti geni sono necessari per la vita, come può essere che un verme abbia più geni di una mosca, ed un uomo solo pochi più di un verme e, naturalmente, possiamo sempre riprendere il discorso sul significato della vita.

# CAPITOLO II

# La tecnologia Microarray

## 2.1 Introduzione

La tecnologia microarray fa uso delle sequenze create dai progetti genoma per cercare quale gene è espresso in un particolare tipo di cellula di un particolare organismo in un istante particolare e sotto particolari condizioni; ad esempio permette il confronto tra espressioni di geni tra cellule normali e malate.

La tecnologia microarray è composta da due parti fondamentali, una parte hardware, formata da macchinari che preparano i vetrini che sono di supporto all'esperimento, da laser e scanner, che rilevano i risultati dell'esperimento e da una parte software formata da programmi che analizzano i risultati dell'esperimento. Vedremo in seguito in cosa consiste un esperimento microarray e che tipo di analisi i bisogna fare sui dati per ottenere informazioni utili ai biologi.

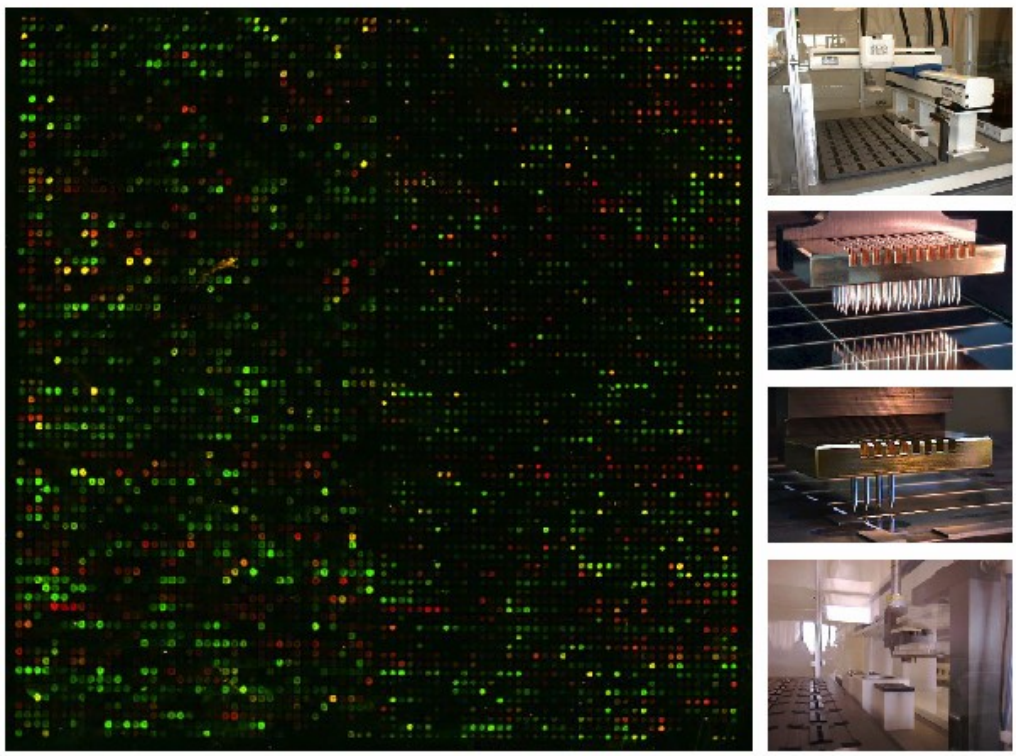


*Figura 12: Immagine del microarray dell'intero genoma del lievito (sinistra); macchinari per la costruzione dei microarray (destra) e testine di stampa dei vetrini (destra al centro).*

E’ importante ribadire che questa tecnologia si basa sulla conoscenza dei geni del DNA degli organismi sottoposti ad analisi e quindi è utilizzabile solo per organismi di cui si conosce il genoma (al momento solo di alcuni organismi vegetali e animali e anche quello dell’uomo).

## 2.2 L’esperimento microarray

Un esperimento microarray mette a confronto l’espressione dei geni di due tipi di cellule, un tipo utilizzato come riferimento ed un tipo da sottoporre a test. Un esperimento quindi inizia con la fase di scelta delle cellule da sottoporre ad analisi, si possono scegliere cellule di due tessuti differenti, cosi da rilevare le differenze tra i geni attivi in un tessuto rispetto a quelli di un altro tessuto, si possono scegliere cellule appartenenti allo stesso tessuto ma di due esemplari differenti uno sano ed uno malato, oppure cellule dello stesso tessuto ma sottoposte a due tipi di ambiente diverso, differenti condizioni di temperatura o di alimentazione, oppure infine  cellule dello stesso tessuto ma prelevati ad intervalli di tempo differenti  per analizzare il ciclo cellulare.

Dopo questa fase si procede all’esperimento vero e proprio illustrato nella figura seguente.

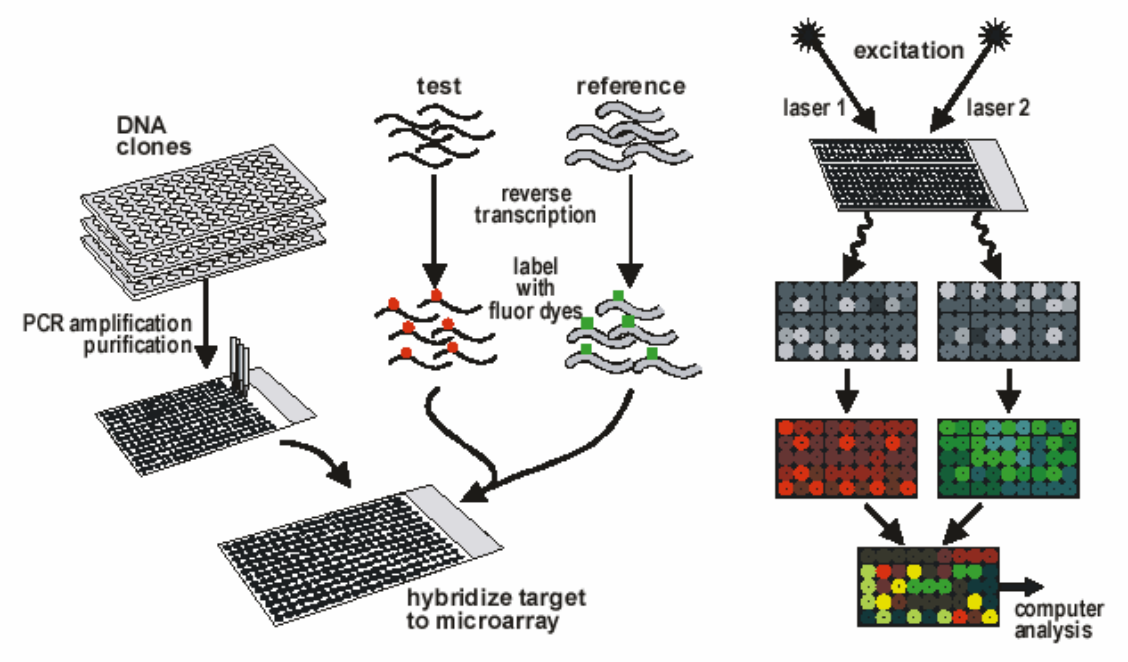



*Figura 13: L'esperimento microarray.*

Dalle cellule scelte viene estratto l'mRNA che indica quali e quanti geni sono espressi in quell'istante, questo è un acido che si degrada velocemente quindi ne viene fatta una trascrizione inversa (reverse transcription) cosi da ottenere un DNA complementare (cDNA). Questo viene colorato con elementi fluorescenti che si legano alle molecole di cDNA e che reagiscono solo alla luce di una specifica lunghezza d'onda cosi da poter essere rilevati con laser di lunghezza d'onda opportuna.

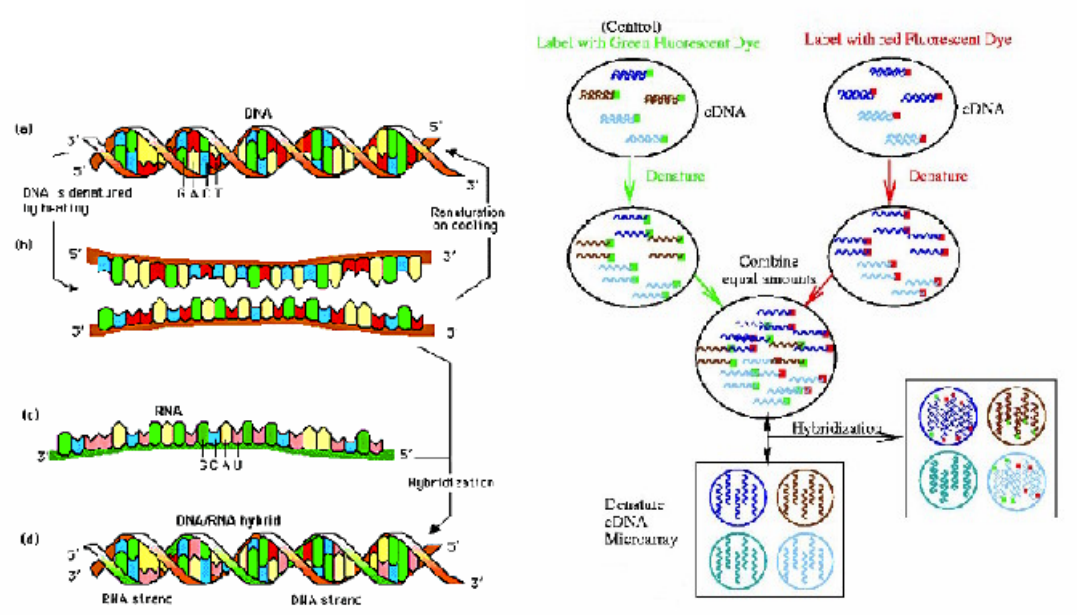



*Figura 14: fase di ibridazione (sinistra); esperimento microarray in dettaglio (destra).*

Finita questa fase inizia la fase di ibridazione: il cDNA ottenuto dei due tipi differenti viene posto in un unico recipiente e poi viene messo a contatto con un vetrino microarray precedentemente preparato. Sul vetrino sono state poste ordinatamente con una struttura a griglia piccole gocce di DNA ognuna di uno specifico gene da analizzare, quindi quando il vetrino è messo in contatto con la miscela dei due cDNA colorati, il DNA del microarray si legherà solo al cDNA dello stesso gene e i colori saranno utilizzati per determinare i livelli di quel gene nei due tipi differenti.

Dopo la fase di ibridazione viene quella di scanning in cui laser a lunghezza d'onda differenti colpiscono il vetrino microarray e tramite un sensore viene

rilevata un’immagine del vetrino, una per ognuno dei due colori. Queste immagini sono in bianco e nero e i colori visualizzati sono solo elaborazioni al computer: alla prima immagine si attribuisce un colore verde alla seconda un colore rosso e unendo le due immagini si ottiene una gradazione di colori dal verde al rosso passando per il giallo. Il colore in questo caso indica i rapporti tra i livelli dei geni nei due tessuti.

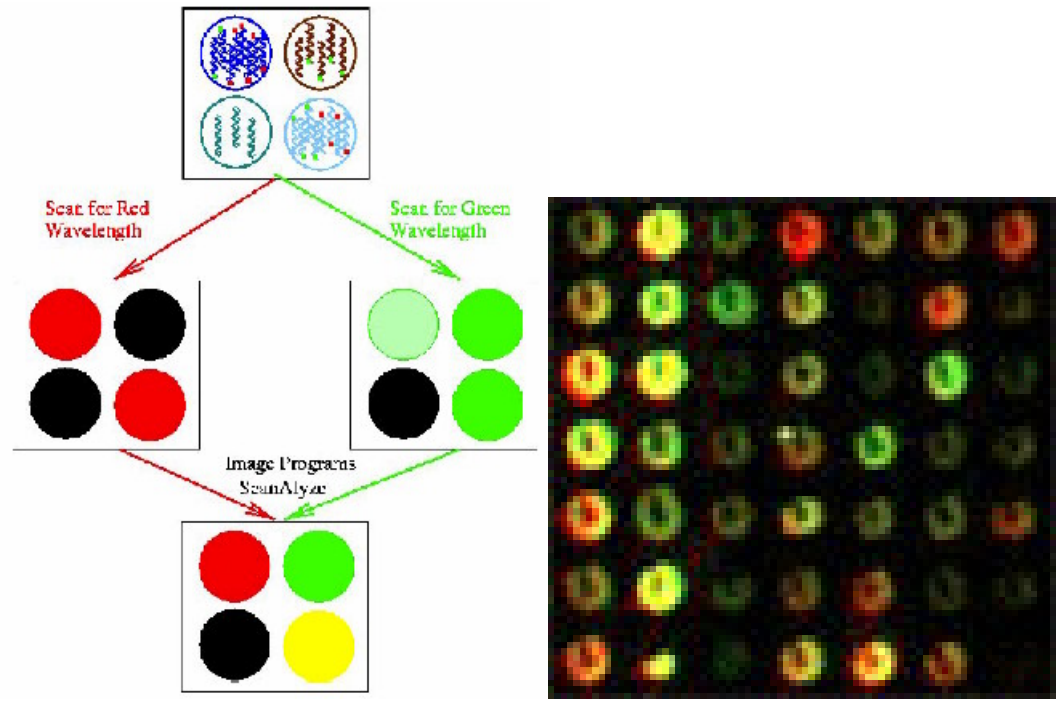



*Figura 15: fase di preparazione dell'immagine da analizzare*

Ottenuta l’immagine si passa dalla parte hardware a quella software e si inizia la fase di analisi dell’immagine.

## 2.3 L’analisi dei dati : Image quantify

Le immagini ottenute dall’esperimento sono sottoposte a vari tipi di analisi il primo tipo di analisi è quello che serve per ottenere dal singolo punto (spot) composto da un insieme di pixel il valore dell’espressione del gene considerato.

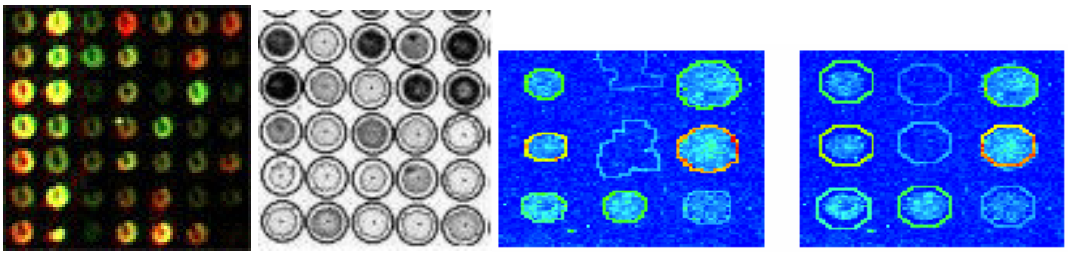


*Figura 16: selezione degli spot per la quantificazione del livello di espressione del gene*

Il punto (spot) viene identificato nella griglia e viene assegnato al gene corrispondente un numero che indica il livello di espressione di quel gene in base ad analisi statistiche dell' immagine dello spot considerato.

Una volta che viene assegnato per ogni colore il livello di espressione del gene, per confrontare i due colori bisogna procedere alla normalizzazione delle due matrici delle intensità. Facendo l'ipotesi che l'insieme dei valori ottenuti dalle due matrici abbia l'andamento di una curva gaussiana si traslano i valori in modo che la media sia uguale a zero per i due colori così che quando si calcolano i rapporti delle intensità questo valore non sia affetto da errori dovuti a differenze di fluorescenza ma sia espressione del rapporto di espressione dei geni.

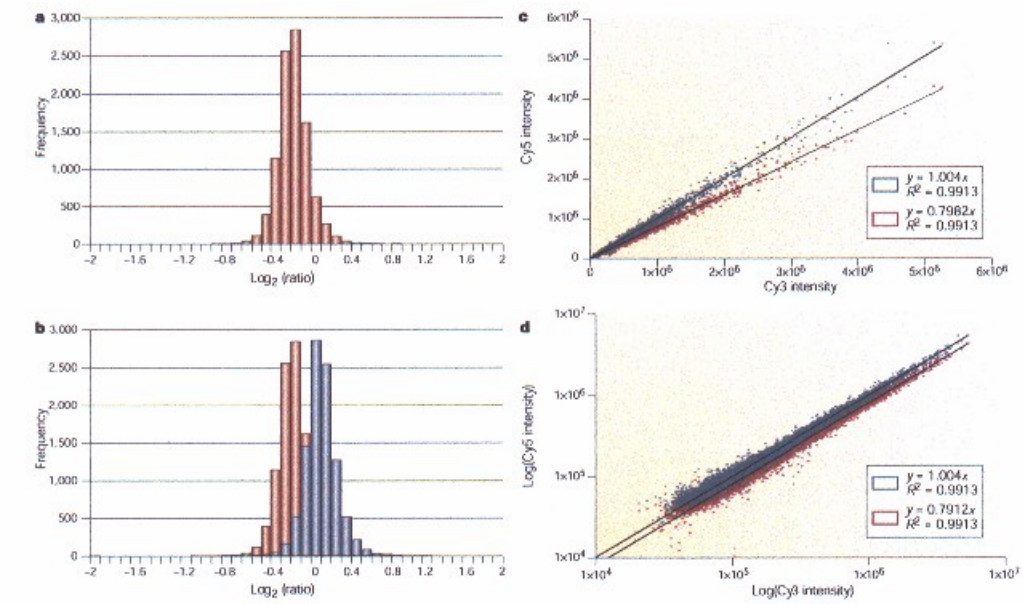



*Figura 17: In rosso ci sono i dati originali in blu quelli normalizzati.*

Il rapporto però crea una inefficienza nella rappresentazione perché i geni sovraespressi avranno un valore del rapporto nell'intervallo $(1,+\infty)$ mentre per i geni sottoespressi si avrà un valore nell'insieme (0,1). Per superare questo fatto i dati sono rappresentati come il logaritmo del rapporto tra le intensità di colore: $y_{ij} = \log_2\left(\frac{Cy5_{ij}}{Cy3_{ij}}\right)$

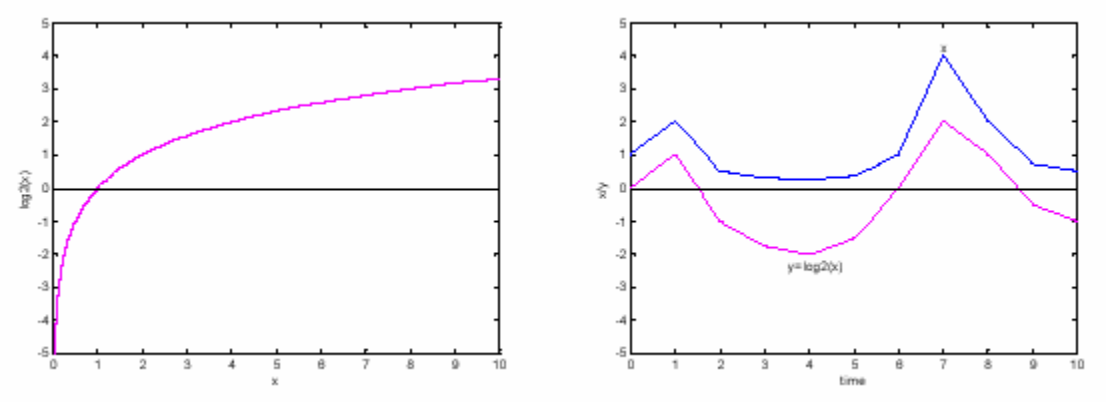



*Figura 18: (sinistra) funzione logaritmo in base 2, (destra) espressione del gene normale (blu) e dopo la trasformazione (magenta)*

Dopo la normalizzazione i dati relativi ad un singolo esperimento sono pronti per essere confrontati con i dati di altri esperimenti ed è questa la vera novità della metodologia. Si costruisce quindi a partire dai vari esperimenti una matrice che ingloba tutti i dati di molti esperimenti relativi ai molti geni sottoposti ad analisi. Si ha quindi una più complessa analisi dei dati.

## 2.4 L'analisi dei dati : algoritmi di normalizzazione e di data mining

Con l'insieme dei dati sui geni relativi a molti esperimenti si costruisce la matrice delle espressioni dei geni in cui le righe rappresentano l'andamento dell'espressione del gene in funzione dell'esperimento e le colonne rappresentano il livello di espressione di ogni gene per il singolo esperimento, il valore di un elemento della matrice ,espresso con un colore, rappresenta il logaritmo del rapporto del livello di espressione del particolare gene nel particolare esperimento per le cellule da testare e quelle di riferimento.

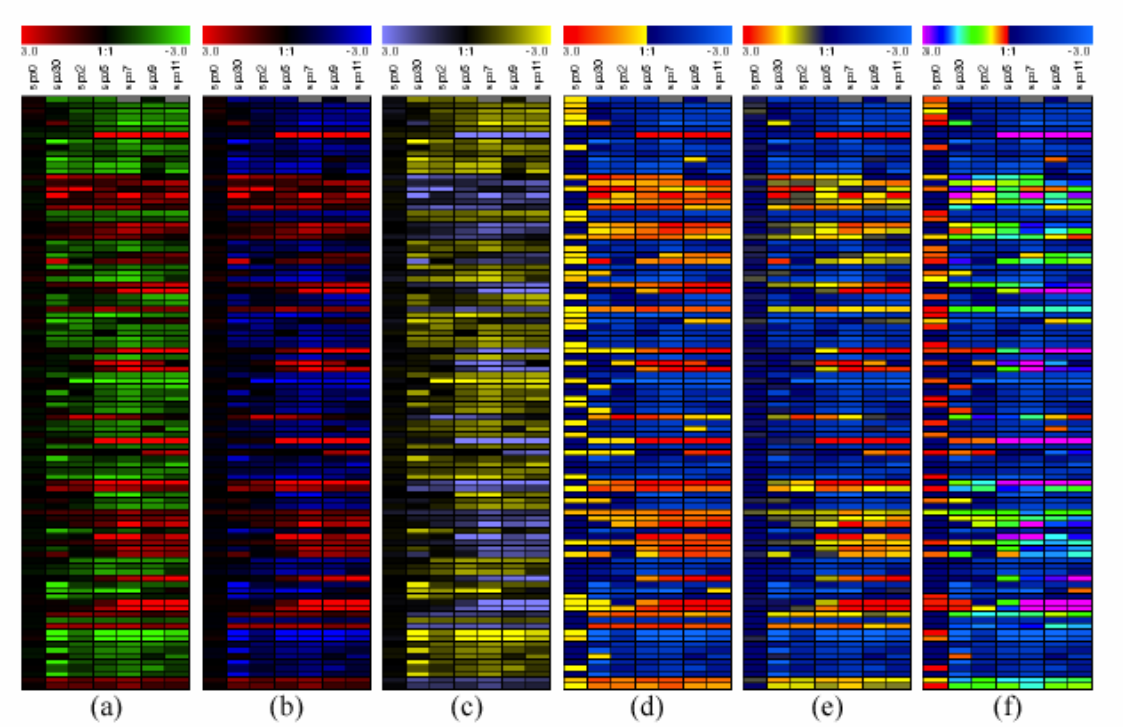



*Figura 19:diversi rappresentazioni grafiche per gli stessi dati: la più usata (a), per persone daltoniche(b), per risaltare gli estremi(c), per differenziare i geni sotto e sovra espressi (d), evidenzia solo i sovra espressi (e), gradiente solo per i sovra espressi (f).*

Nella seguente immagine viene mostrato un modo alternativo di rappresentare i dati in cui le espressioni dei geni sono funzione degli esperimenti condotti ad istanti di tempo successivi, la spezzata in magenta è la media di tutte le espressioni di tutti i geni

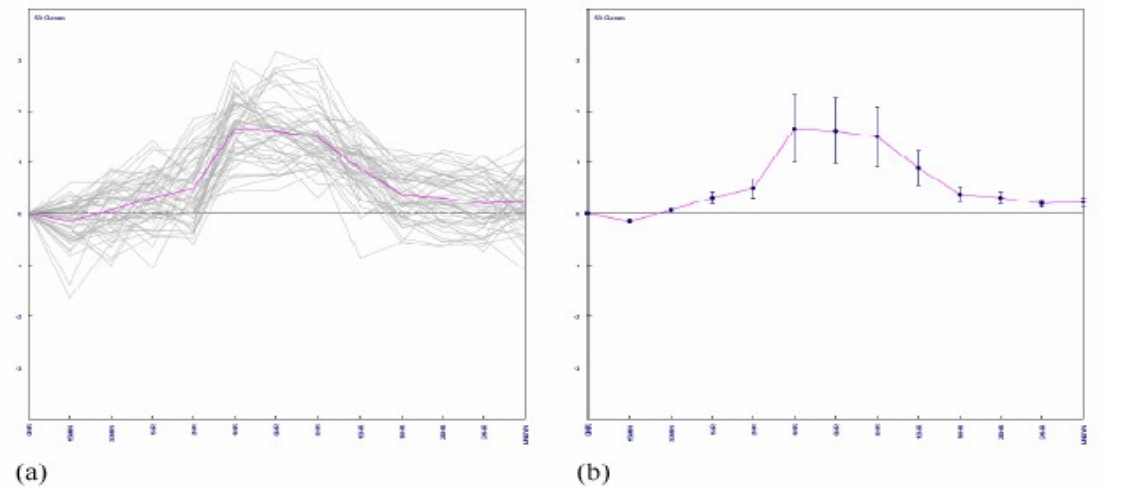



*Figura 20: grafico delle espressioni dei geni: (a) sono disegnati tutti i geni e la curva viola è la media delle curve di tutti i geni; (b) media e deviazione standard normalizzata dei geni in (a)*

Prima dell'analisi statistica dei dati sono spesso usate procedure per la normalizzazione degli stessi, le normalizzazioni più usate sono la sottrazione della media dei valori da una espressione o (mean centering) e la divisione dei valori dell'espressione per il valore efficace (o valore RMS) o per la varianza.
La prima di queste trasformazioni rende confrontabile espressioni di geni che si riferiscono ad esperimenti diversi,in questo modo si confrontano le variazioni delle espressioni dei geni come è illustrato nella figura seguente.

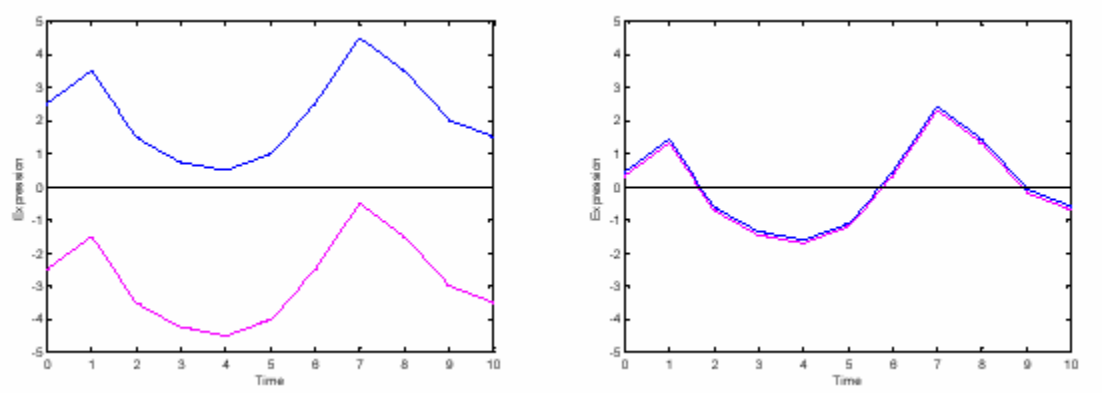



*Figura 21: (sinistra) espressioni di due geni uno sovraespresso e l'altro sottoespresso, (destra) dopo la normalizzazione la distanza tra loro è nulla.*

Questo tipo di normalizzazione è usato anche per eliminare per un singolo esperimento dominanti di colore dovuti al procedimento tecnologico non esente da questo tipo di errori.
Un altro tipo di normalizzazione e quello della divisione dei dati per il valore efficace o la per la varianza, questo tipo di trasformazione in una espressione amplifica i valori piccoli e riduce quelli grandi in modulo. Ha l'inconveniente di amplificare anche il rumore rendendo le espressioni simili.

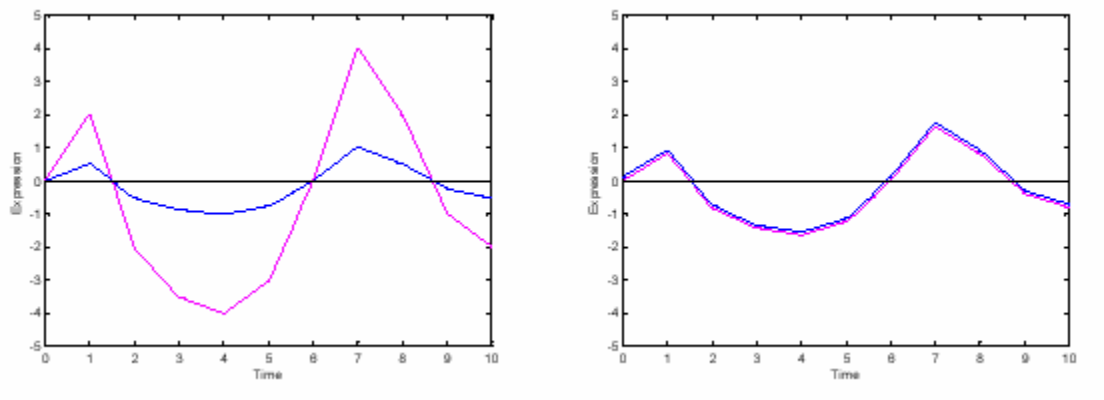



*Figura 22: divisione per RMS o DS, segnali non elaborati (sinistra), i segnali elaborati possono essere molto simili (destra)*

Lo scopo dei vari algoritmi di data mining utilizzati è quello di raggruppare le espressioni dei geni simili, questo è molto importante dal punto di vista biologico perché indica che o quei geni possono partecipare ad uno stesso processo biologico o che sono regolati dagli stessi meccanismi, questo aiuta a scoprire il funzionamento della cellula e le interazioni tra i geni, tra i geni e le proteine e le interazioni tra la cellula e l'ambiente esterno.

Il concetto di similitudine o di vicinanza è legato alla metrica utilizzata per effettuare la misura della distanza tra le righe della matrice, le espressioni dei geni, che in questo caso vengono interpretate come vettori. I programmi che implementano questi algoritmi quindi mettono a disposizione anche un gran numero di metriche.

Un tipo di raggruppamento (clustering) molto utilizzato è il clustering gerarchico, che raggruppa le espressioni simili partendo da quelle più vicine per poi aggregare a questo gruppo altre espressioni secondo uno dei tre metodi di collegamento implementati finora. Nel collegamento singolo (single linkage) la distanza tra due cluster è determinata dalla distanza dei due elementi più vicini, nel collegamento completo (complete linkage) la distanza tra cluster è determinata dalla distanza degli elementi più lontani, mentre nel collegamento medio (average linkage) le distanze tra i cluster e determinata a partire dalla media degli elementi in ciascun cluster.

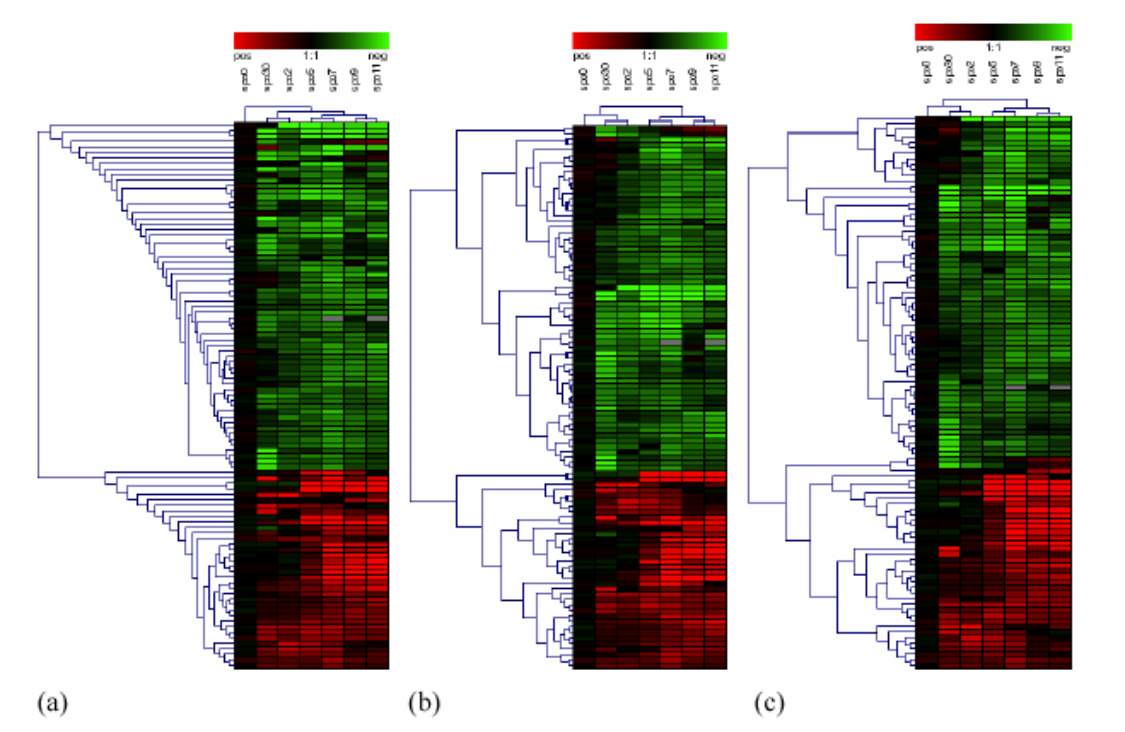



*Figura 23:esempio di clustering gerarchico con diversi metodi di raggruppamento (linkage): singolo(a), completo(b), medio(c).*

Un altro importante tipo di analisi è quello di utilizzare le reti neurali ed in particolare le reti SOM (Self-Organizing Maps) per individuare i gruppi (cluster) di geni con le stesse espressioni.

In uno spazio a più dimensioni si distribuiscono in modo regolare dei nodi sonda che dovranno individuare i cluster, per ogni gene si calcola la distanza del gene con ogni nodo e quello più vicino viene avvicinato al gene gli altri no oppure vengono allontanati, in questo modo dopo che tutti i geni sono stati inseriti i nodi sonda saranno vicino al centro del cluster.

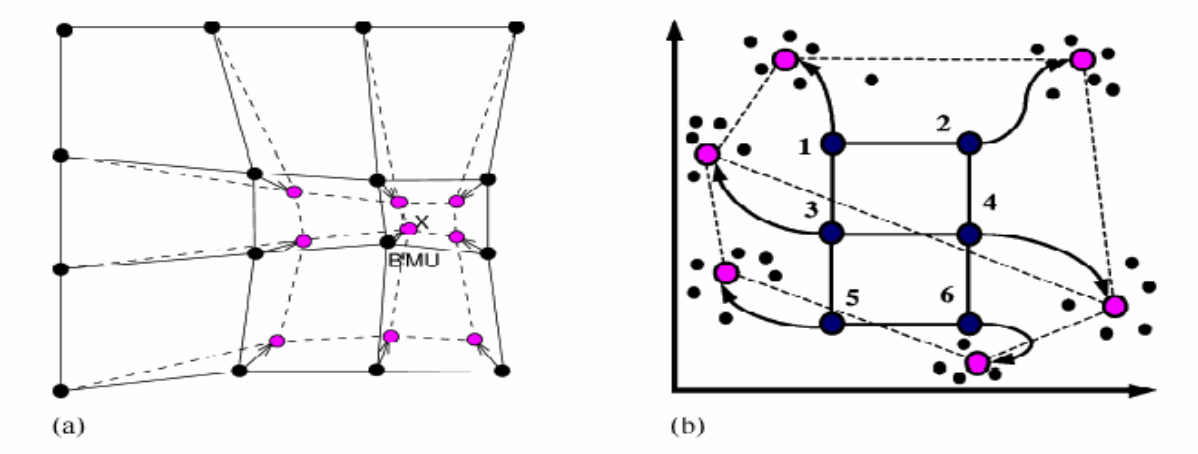



*Figura 24: Modificazione della rete dopo una interazione (a), alla fine delle interazioni (b).*

Continuando la panoramica degli algoritmi di analisi dobbiamo ricordare la K-mean analisis, il raggruppamento parte dalla decisione del numero di gruppi (cluster) in cui dividere i geni, questa scelta è fatta empiricamente dal ricercatore, poi i vettori rappresentanti le espressioni dei geni vengono suddivisi tra in modo casuale nei vari cluster e viene calcolato il vettore medio di ogni cluster, infine con un processo iterativo viene calcolata la distanza di ogni vettore dal vettore medio di ogni cluster e il vettore viene spostato nel cluster da cui ha la minima distanza, quando tutti i vettori sono stati processati si ricalcolano le medie per ogni cluster e il procedimento si itera fino a che non ci sono più spostamenti di vettori.

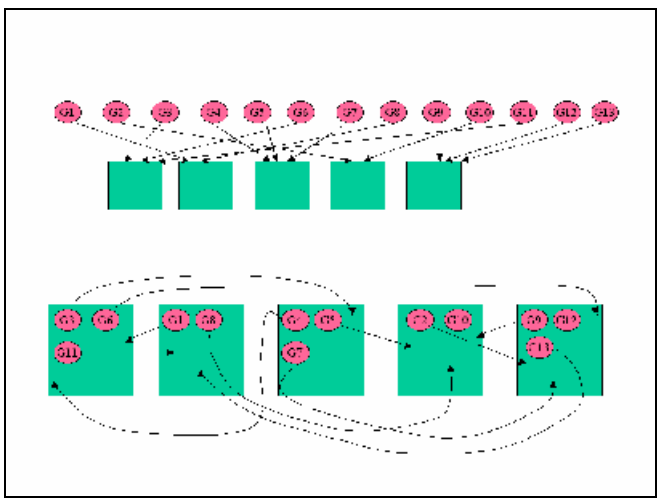


*Figura 25: K-means clustering.*

L'ultimo tipo di analisi non supervisionata che illustriamo è la *principal component analisis* (PCA), questo tipo di analisi ha lo scopo di ridurre la complessità della rappresentazione dei geni. Supponiamo che abbiamo la matrice delle espressioni dei geni, e che le espressioni di alcuni di essi siano correlate, la PCA ignora gli esperimenti ridondanti e fornisce una media pesata di alcuni esperimenti così da fornire un andamento dei dati più interpretabile. I componenti possono essere pensati come assi in uno spazio n-dimensionale, dove n è il numero di componenti, ogni asse rappresenta un andamento differente nei dati. Nelle figure seguenti è mostrato un esempio di visualizzazione dei geni mediante la PCA.

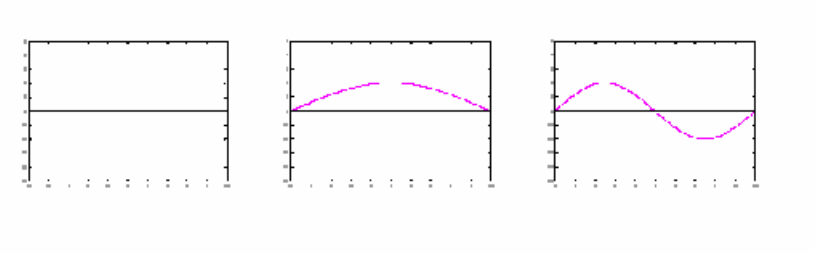


*Figura 26: Indichiamo da sinistra a destra con PCA1,PCA2 ePCA3 i componenti principali individuati*

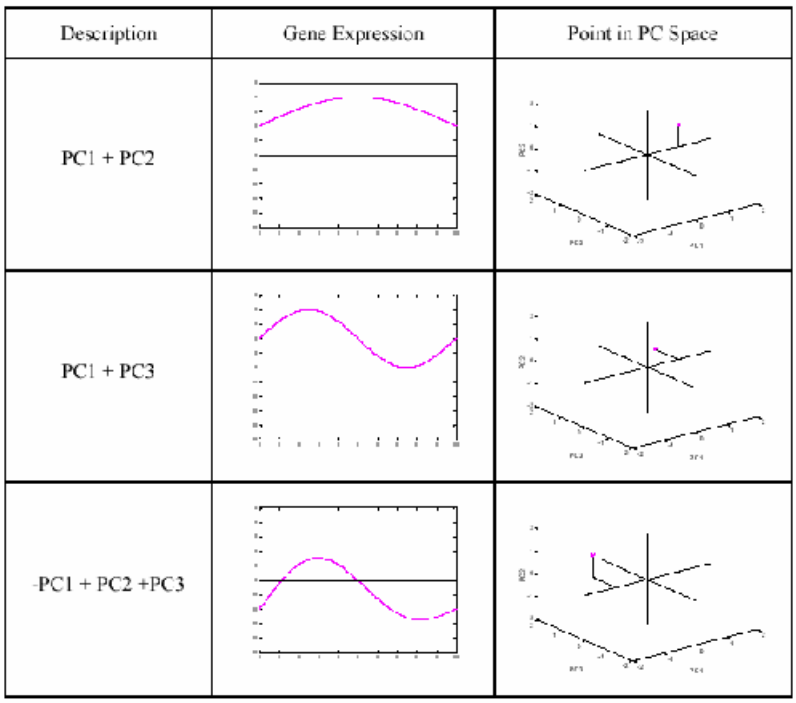



*Figura 27: Le espressioni dei geni possono essere visualizzate come punti dello spazio a 3 dimensioni come combinazione lineare dei tre componenti principali individuati in fig.27.*

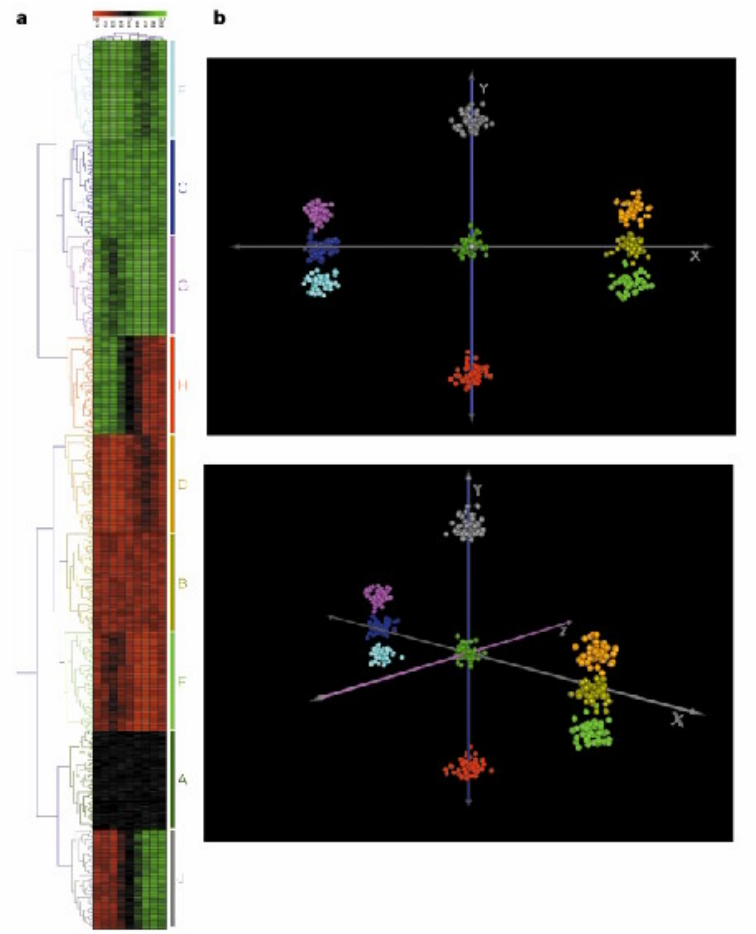



*Figura 28: confronto tra cluster analisis (a) e PCA (b) si può notare come sia immediato identificare i diversi cluster nei grafi 3D della PCA.*

Accanto ai metodi di analisi non supervisionati ci sono metodi di analisi supervisionati in cui si hanno a disposizione informazioni sui cluster e quindi si vuole decidere se un gene appartiene o meno a quel determinato cluster. Il metodo più usato è il support vector machine (SVM), sostanzialmente una rete neurale che usa un training set di geni di cui alcuni sono relazionati tra loro, ad esempio sono attivati per una stessa funzione biologica, e sono indicati come esempi positivi di appartenenza ad un cluster ed altri non collegati con i primi che fungono da esempi negativi. Il training set è utilizzato dalla SVM per apprendere quali geni fanno parte della classe in esame e quali no. Dopo questa fase la macchina è pronta a classificare i geni sulla base della loro espressione. In generale quindi la SVM usa informazioni biologiche per identificare un

gruppo di geni e per decidere se i geni in analisi fanno parte o meno di quel gruppo.

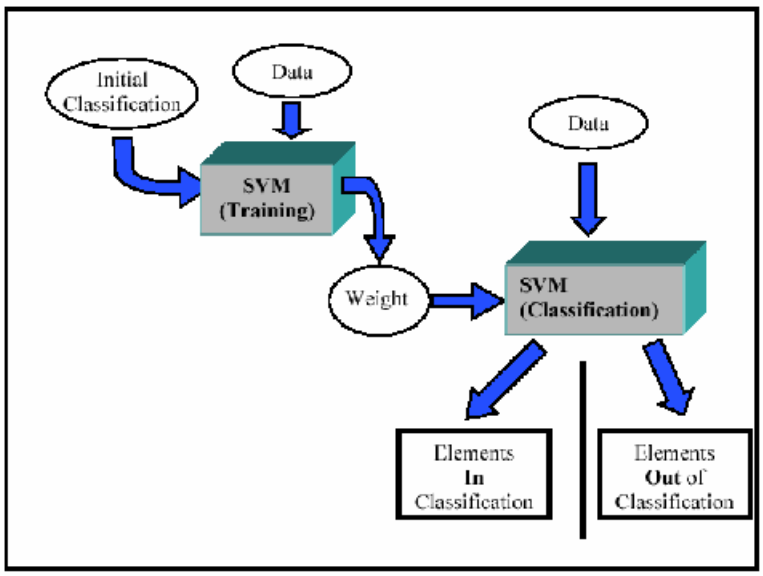



*Figura 29: Funzionamento della SVM: nella prima fase (training) la macchina apprende, nella seconda (classification) utilizza l'apprendimento per discriminare i dati in analisi.*

## 2.5 L'analisi dei dati: la visualizzazione delle reti di interazione tra i geni

Il passo successivo nell'analisi è quello di inserire i risultati delle analisi precedenti in un contesto più ampio che comprenda i processi biologici che si vogliono studiare e i dati relativi ad altri prodotti di questi processi come le proteine sintetizzate e le sostanze che intervengono nel processo.

I programmi di ausilio a questa analisi mostrano le reti di interazioni tra i geni di vari processi biologici già conosciuti oppure permettono di creare nuove reti da valutare dal punto di vista biologico.

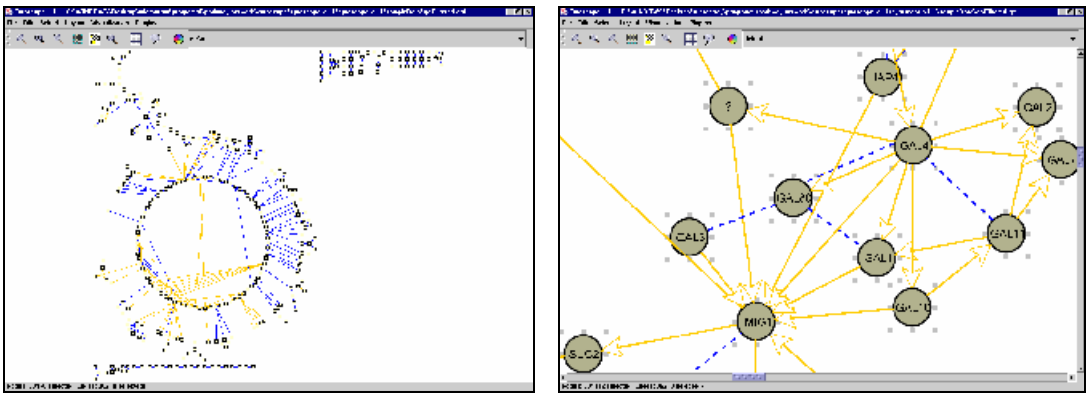


*Figura 30: rete costruita a partire dalle relazioni individuate con analisi precedenti (sinistra), particolare (destra).*

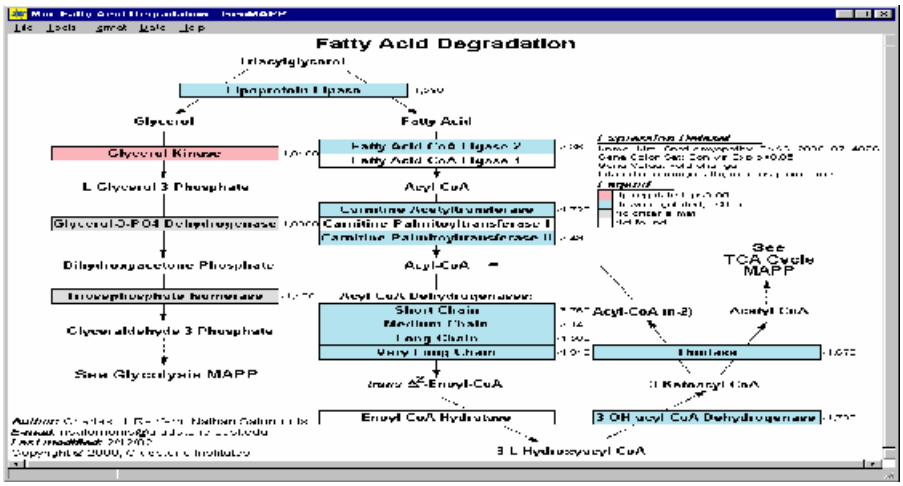



*Figura 31: rete di interazioni gia conosciute in cui i colori indicano le espressioni dei geni analizzate*

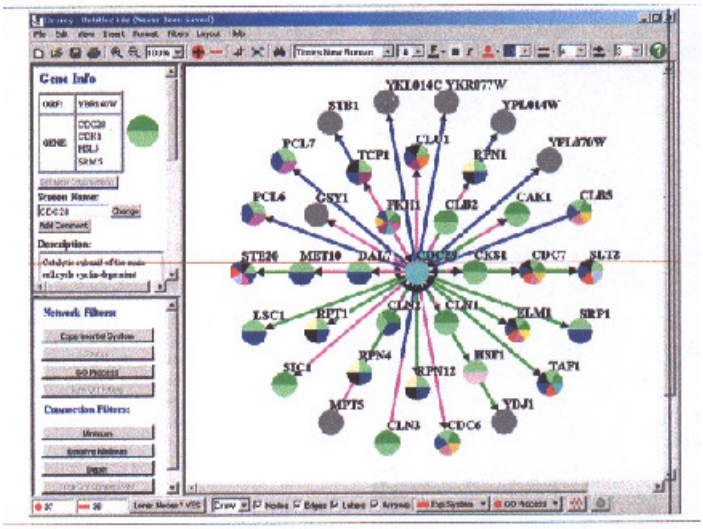



*Figura 32: In questa rappresentazione le caratteristiche dei geni (nodi) e delle interazioni (rami) sono codificate con uno schema di colori particolare.*

# CAPITOLO III

# Software per l' Image Quantify

## 3.1 Introduzione

I software per l'image quantify forniscono tre funzioni fondamentali per il raggiungimento dei risultati, il gridding, che è il processo di individuazione dei singoli punti (spots) nell'immagine, la segmentazione, che è il processo che separa i pixel appartenenti allo spot da quelli appartenenti allo sfondo e infine l'estrazione delle informazioni, che si divide in due parti, l'estrazione di informazione dai punti e l'estrazione di informazioni dallo sfondo.

Le informazioni estratte sono le caratteristiche dello spot come l'area dello spot, la sua forma, la luminosità media, la varianza delle luminosità ed altri dati statistici per meglio individuare il livello di espressione del gene dello spot.

## 3.2 Dapple: Image Analysis Software for DNA Microarrays.

**Autori**

Jeremy Buhler lab;

**Organizzazione**

Department of Computer Science, Washington University in St. Louis; St. Louis, Missouri – USA.

**Sistema operativo e ambiente operativo**

Sistemi operativi Unix-like (testato su vari sistemi unix: linux, AIX, Solaris, Tru64).

**Versione**

0.87

**Licenza**

GNU General Public License.

**Note**

Il software è fornito in formato sorgente, in linguaggio C++; sono fornite le istruzioni di compilazione e il manuale di utilizzo.

**Sito web:**

http://www.cs.wustl.edu/%7Ejbuhler//research/dapple/

**Funzionamento**

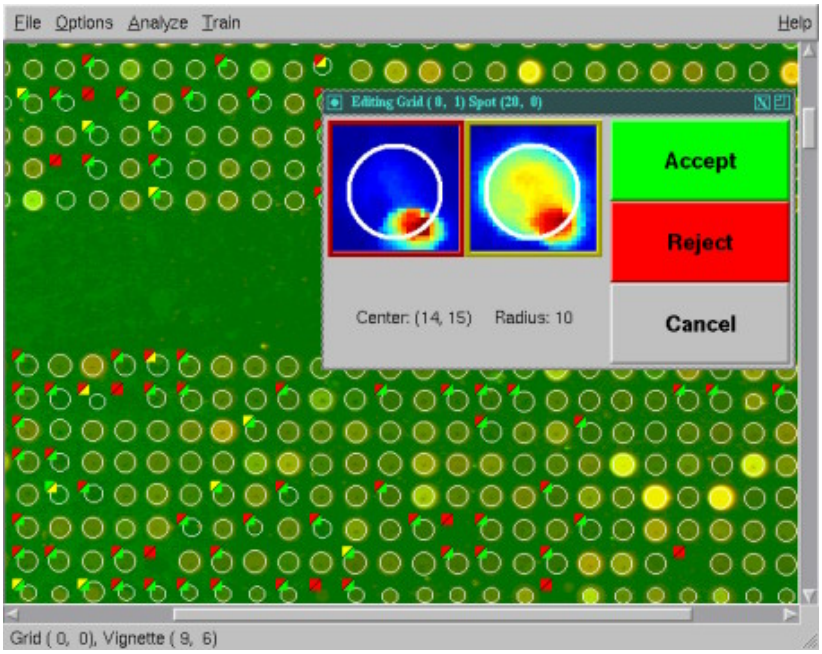
Dapple è un programma per la quantificazione dei punti (spots) in un'immagine microarray.

Data una coppia di immagini (in formato .TIFF) ognuna trattata con un diverso agente fluorescente (rosso e verde) dapple cerca i singoli punti nell'immagine unione delle due in ingresso, valuta la loro qualità (forma, contorni, ...) e quantifica la fluorescenza totale di ogni punto.

Il programma è progettato per individuare i punti anche in presenza di irregolarità della loro grandezza o posizione dovute alle vibrazioni del robot che crea i microarray o ad altre cause di rumore sulle immagini.

Il programma giudica la qualità del punto automaticamente e lascia quelli dubbi al giudizio dell'utente.

**Screenshot:**


File Options Analyze Train
Help
Editing Grid ( 0, 1) Spot (20, 0)
Accept
Reject
Cancel
Center: (14, 15) Radius: 10
Grid ( 0, 0), Vignette ( 9, 6)

## 3.3 F-scan

**Autori**

P. J. Munson, V. V. Prabhu, L. Young;

**Organizzazione**

Analytical Biostatistics Section, Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health; Rockville, Maryland,USA.

**Sistema operativo e/o ambiente operativo**

MATLAB 5.2 o superiore, indipendentemente dal sistema operativo adottato.

**Versione**

1.3

**Licenza**

Gratuito per uso accademico (non commerciale).

**Note**

Il software è fornito in formato sorgente, in linguaggio MATLAB; bisogna registrarsi per poter accedere ai programmi; è fornito il manuale di utilizzo.

**Sito web:**

http://abs.cit.nih.gov/fscan/

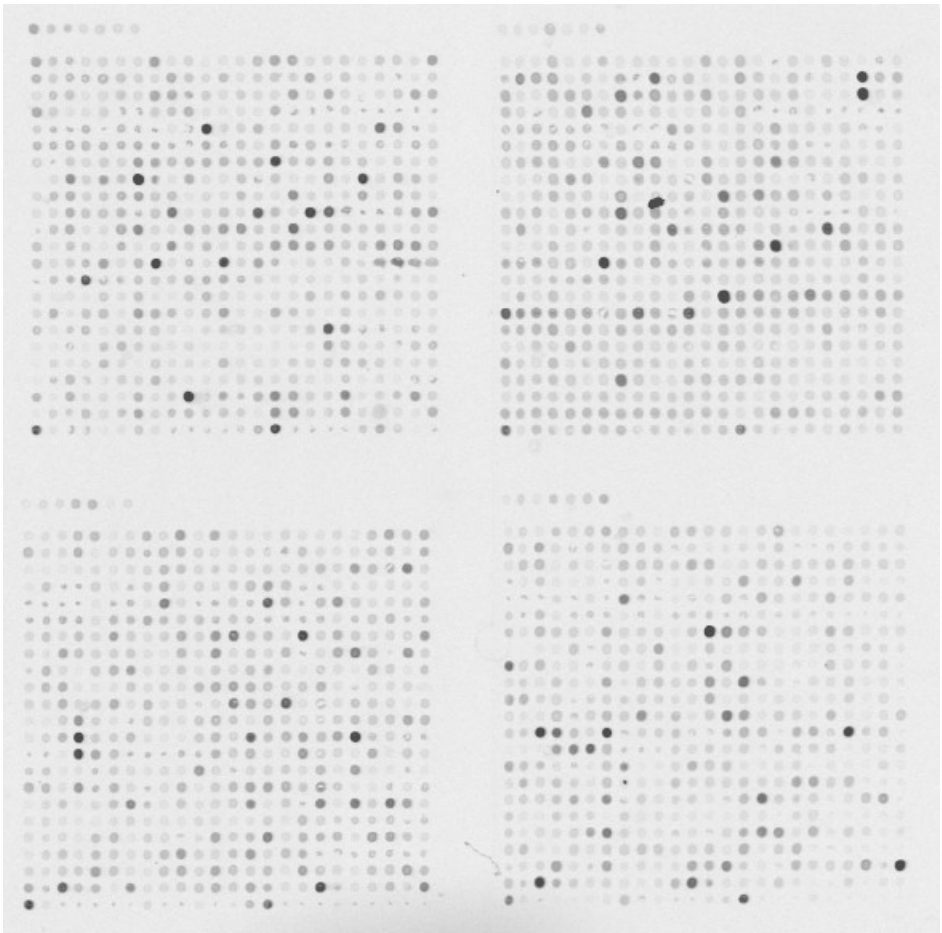
**Funzionamento:**

Da una coppia di immagini, genera un file contenente la posizione e l'intensità dei punti e visualizza un immagine dei punti.
Compara le intensità dei punti tra il canale rosso e verde delle immagini e genera una lista di geni sotto e sovra espressi.

Per analisi statistiche sofisticate prepara un file contenente le intensità di tutte le immagini che può essere aperto da software statistici come JMP. Si possono analizzare una serie di esperimenti. E' mostrata a video l'immagine composta dalle due in input.

Inputs :

I file delle due immagini (una per ogni canale) del microarray da analizzare in formato .gel :



La lista delle corrispondenze gene-punto del microarray(formato .gal):

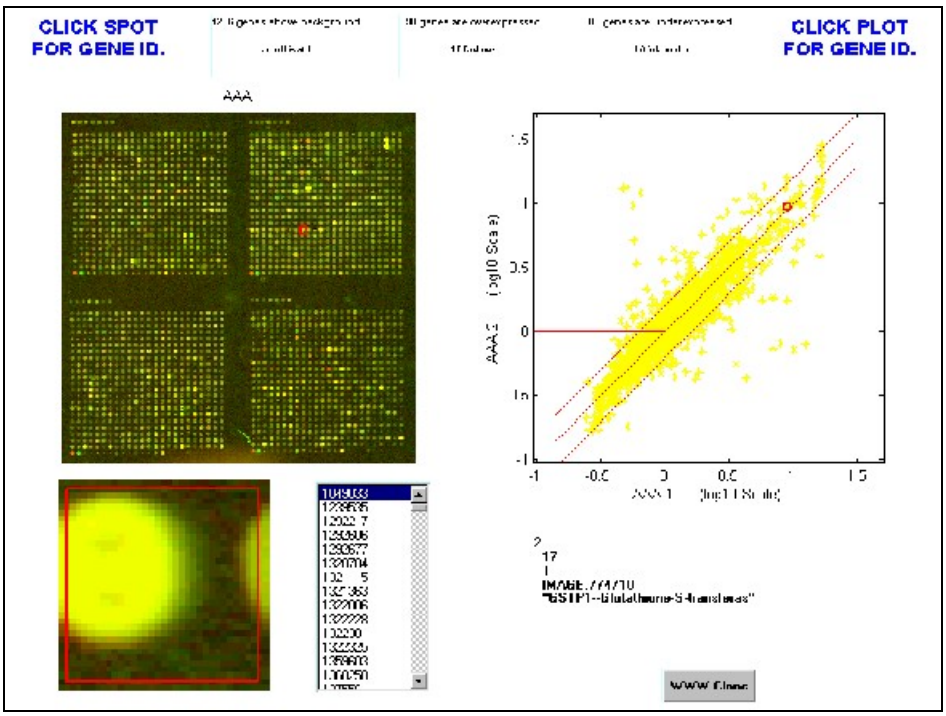
```
ATF     1.0
7       5
"Type=GenePix ArrayList V1.0"
"BlockCount=4"
"BlockType=0"
"Block1=        2750    22900   220     23      325     24      325"
"Block2=        11700   22900   220     23      325     24      325"
"Block3=        2750    31860   220     23      325     24      325"
"Block4=        11700   31860   220     23      325     24      325"
"Block" "Row"   "Column"        "ID"    "Name"
```

```
1       1       1       IMAGE:812965    "MYC--c-myc"
1       1       2       IMAGE:260303    "ETS2--ets-2=ets family transcri"
1       1       3       IMAGE:526282    "CSK--csk=C-SRC-kinase"
1       1       4       IMAGE:724588    "ISGF3G--ISGF3 gamma=IFN alpha/b"
1       1       5       IMAGE:246300    "TIAL1--TIAR=nucleolysin=cytotox"
1       1       6       IMAGE:768561    "SCYA2--MCP-1=MCAF=small inducib"
1       1       7       IMAGE:813256    "ABCB1--MDR1=Multidrug resistanc"
1       1       8       IMAGE:811900    "LTBR--Lymphotoxin-Beta receptor"
```

Output:

Un file in formato .mat (matrice in matlab) e un file in formato .pGL per le analisi statistiche.

**Screenshot:**



## 3.4 P-scan (Peak quantification using Statistical Comparative Analysis)

**Autori**

P. J. Munson, V. V. Prabhu, L. Young;

**Organizzazione**

Analytical Biostatistics Section, Mathematical and Statistical Computing Laboratory; Center for Information Technology, National Institutes of Health; Rockville, Maryland, USA.

**Sistema operativo e/o ambiente operativo**

MATLAB 5.2 o superiore, indipendentemente dal sistema operativo adottato.

**Versione**

1.2

**Licenza**

Gratuito per uso accademico (uso non commerciale).

**Note**

Il software è fornito in formato sorgente, in linguaggio MATLAB,previa registrazione ; è fornito il manuale di utilizzo.

**Sito web**

http://abs.cit.nih.gov/pscan/index.html

**Funzionamento**

Per ogni immagine genera un file contenente la posizione e l'intensità dei punti. Compara le intensità dei punti tra il canale rosso e verde delle immagini e genera una lista di geni sotto e sovra espressi.

Per analisi statistiche sofisticate prepara un file contenente le intensità di tutte le immagini che può essere aperto da software statistici come JMP.

Inputs:

I file delle due immagini (una per ogni canale) del microarray da analizzare in formato .gel e la lista delle corrispondenze gene-punto del microarray:

| ROW | COLUMN | Gene Name | GenBank Accession #(s) | Atlas Human Cancer Array (#7742 | Atlas Human Oncogene/ Tumor Sup | SPOT | KIND | 33-row |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | cell division control protein 2 homolog (EC 2.7.1.-); P34 protein kinase; cyclin-dependent kinase 1 (CDK1) | X05360 | A1a | | 0 | 32 | |
| 1 | 3 | CLK-3 | L29220 | A2a | 10F | 0 | 32 | |
| 1 | 5 | cyclin E | M73812 | A3a | 9B | 0 | 32 | |
| 1 | 7 | CDC27HS Protein | U00001 | A4a | | 0 | 32 | |
| 1 | 9 | dual specificity mitogen-activated protein kinase kinase 5 (MAP kinase kinase 5; MAPKK 5) | U25265 | A5a | | 0 | 32 | |
| 1 | 11 | growth inhibitor p33ING1 (ING1) | AF001954 | A6a | 15E | 0 | 32 | |
| 1 | 13 | type I cytoskeletal 16 keratin; cytokeratin 16 (K16; CK 16); pseudo-keratin K16 type I | M21772; M20336 | A7a | | 0 | 32 | |
| 1 | 15 | | | | | 2 | 32 | |
| 1 | 17 | apoptosis regulator bcl-2 | M14745 | B1a | 10B | 0 | 32 | |
| 1 | 19 | receptor interacting protein; serine/threonine protein kinase RIP transferase; serine/threonine-protein kinase ATP-binding apoptosis; cell death protein RIP | U25994; U50062 | B2a | 9D | 0 | 32 | |
| 1 | 21 | interleukin-1 beta convertase precursor (IL-1BC); IL-1 beta converting enzyme (ICE); p45; caspase-1 (CASP-1) | U13699; M87507; X65019 | B3a | | 0 | 32 | |
| 1 | 23 | FAS/APO 1 | Z70519 | B4a | | 0 | 32 | |
| 1 | 25 | calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase 1B (CAM-PDE1B); HCAM-2 | U56976 | B5a | | 0 | 32 | |
| 1 | 27 | C-fos | K00650 | B6a | 19B | 0 | 32 | |

Outputs:

Un file in formato .mat (matrice in matlab) e un file in formato .pGL per le analisi statistiche.

## 3.5 GridGrinder

**Autori**

Tanner C., Tepesch P., and Shulyakov M.;

**Organizzazione**

SourceForge,  Corning Inc.

**Sistema operativo e/o ambiente operativo**

Windows 95/98/NT

**Versione**

1.3

**Licenza**

BSD license (open-source).

**Note**

E’ fornito il programma eseguibile, il manuale utente e il codice sorgente della interfaccia grafica (GUI) e del motore di ricerca dei punti (Engine) , che sono le due parti in cui è diviso il programma.

Il motore è scritto in ANSI C mentre la GUI è scritta in C++ con le MFC (librerie grafiche) di Microsoft.

**Sito web:**

http://gridgrinder.sourceforge.net/

**Funzionamento**

Il programma offre un analisi delle immagini individuando vari tipi di griglia e anche griglie anomale di punti. Per ogni punto individua il centro il contorno il raggio presunto e le intensità.

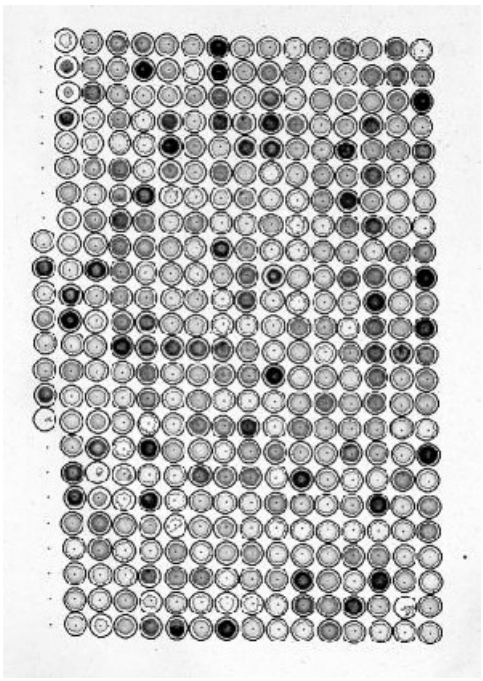
Il programma permette anche la scelta del formato e del contenuto dei dati di output.

Inputs:

I file delle due immagini del microarray da analizzare in formato .tiff :

Outputs:

Un file dell'immagine del microarray elaborato in formato .tif o .jpg:



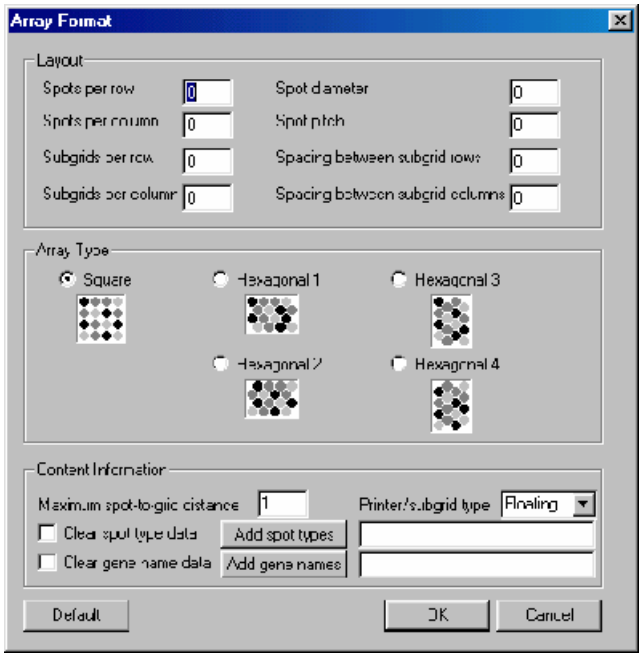
Un file di testo con le informazioni scelte:

#GridGrinder version 0.778
#Number of images in analysis 1
#Reference image TRUE
#Reference image filename C:\WINDOWS\Desktop\microarray\programmi\gridgrinder\16x24-Cy3.tif
#Analyzed image filename C:\WINDOWS\Desktop\microarray\programmi\gridgrinder\16x24-Cy5.tif

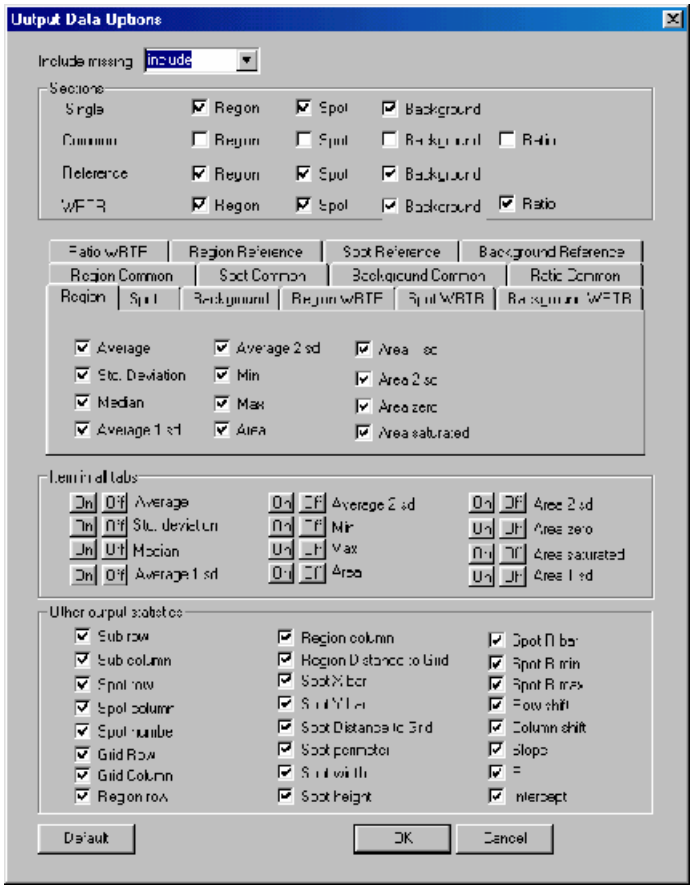
| #SubGrid format | Rectangular | | | Region | | Spot | Background | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SubRow | SubCol | SpotRow | SpotCol | SpotNum | SpotType | Grid Row | Grid Col | Row | Col | Dist-to-Grid |
| | Average | Std_Dev | Median | Avg-1-SD | Avg-2-SD | Area | Area-1-SD | | Area-2-SD | Area- |
| Zero | Area-Satr | Min | Max | Row | Col | Dist-to-Grid | | Average | Std_Dev | Median |
| Avg-1- | SD | Avg-2-SD | Area | Area-1-SD | | Area-2-SD | | Area-Zero | | Area-Satr |
| Min | Max | | Perimeter | Width | Height | Avg Radius | | Min Radius | | Max Radius |
| Average | Std_Dev | | Median | Avg-1-SD | Avg-2-SD | Area | | Area-1-SD | | Area-2-SD |
| Area-Zero | | | Area-Satr | Min | Max | | | | | |
| 1 | 1 | 1 | 1 | 0 | MISSING | 50.000 | 52.000 | 50 | 52 | 0.000 |
| 2654 | | | | | | | | | | |
| 1274 | 2489 | 2504 | 2528 | 797 | 670 | 779 | 0 | 0 | 467 | 16701 |
| 60 | 50 | 10.969 | 3322 | 677 | 3041 | 3088 | 3212 | 38 | 30 | 36 |
| 0 | 0 | 2233 | 5721 | 38 | 11 | 10 | 3.148 | 0.790 | 5.335 | 2556 |
| 1105 | 2367 | 2480 | 2404 | 124 | 91 | 118 | 0 | 0 | 807 | 6682 |
| 1 | 1 | 1 | 2 | 1 | REGULAR | | 50.859 | 87.119 | 49 | 86 |
| 2.170 | | | | | | | | | | |
| 3645 | 2027 | 3148 | 3271 | 3373 | 797 | 630 | 766 | 0 | 0 | 815 |
| 15917 | 48 | 86 | 2.272 | 5659 | 2057 | 5211 | 5164 | 5254 | 267 | 226 |
| 251 | 0 | 0 | 2867 | 15917 | 65 | 20 | 22 | 8.732 | 5.609 | 11.595 |
| 2652 | 1054 | 2493 | 2618 | 2526 | 124 | 90 | 119 | 0 | 0 | 648 |
| 6373 | | | | | | | | | | |

**Screenshots:**

Input dei dati sulla griglia:




Scelta dei dati da fornire nel file di output:



## 3.6 ScanAlyze 2

**Autore**

Michael Eisen;

**Organizzazione**

Michael Eisen's lab, Lawrence Berkeley National Lab (LBNL); Berkeley ,California – USA.

**Sistema operativo e ambiente operativo**

Windows 95/98/NT

**Versione**

2.44

**Licenza**

Gratuita per uso accademico.

**Note**

Il software è fornito in formato eseguibile e sorgente (C++);è fornito il manuale di utilizzo.

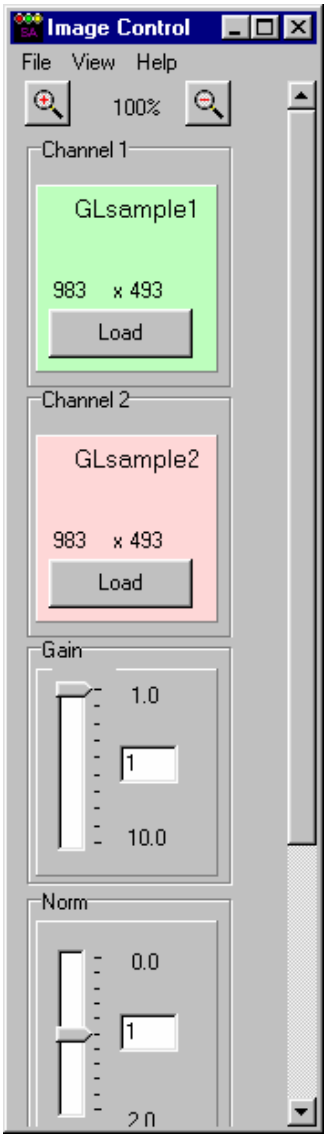
**Sito web:**

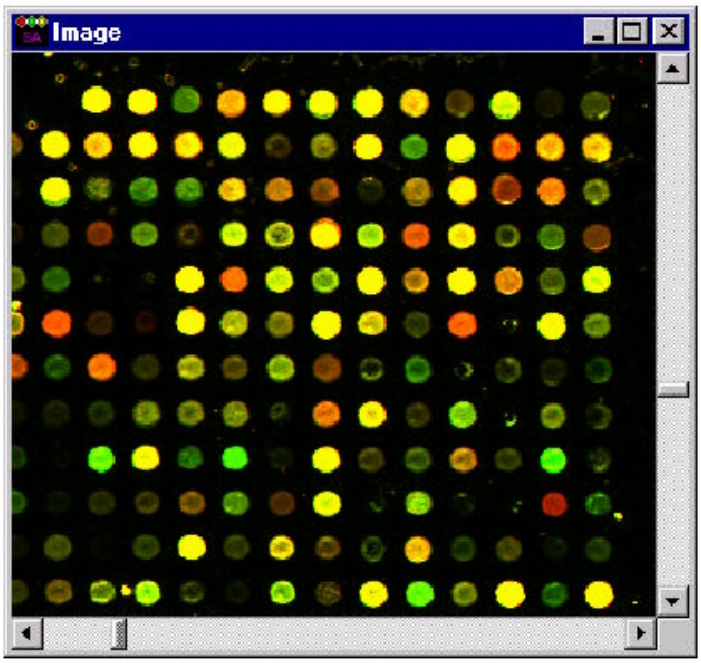
http://rana.lbl.gov/EisenSoftware.htm

**Funzionamento**

Il programma a partire dalla coppia di immagini date dallo scanner effettua una definizione della griglia degli spot in modo semiautomatico ed effettua l'analisi degli spot. I risultati sono riportati in un file in formato testo delimitato dai tab.

**Screenshots:**



## 3.7 TIGR Spotfinder

**Autori**

Vasily Sharov, John Quackenbush;

**Organizzazione**

The Institute of Genomic Research (TIGR); Rockville, Maryland - USA

**Sistema operativo e ambiente operativo**

Windows 95/98/NT

**Versione**

2.0.4

**Licenza**

Gratuita (open source).

**Note**

Il software è fornito in formato eseguibile e sorgente(C++);è fornito il manuale di utilizzo.

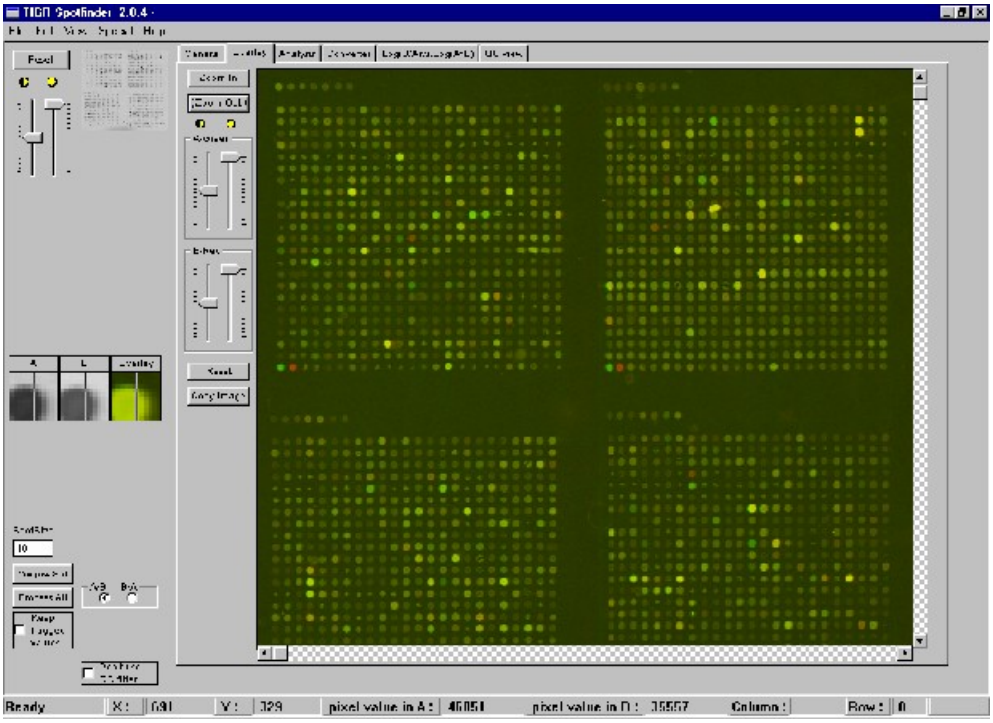
**Sito web:**

http://www.tigr.org/software/tm4/spotfinder.html

**Funzionamento**

Il programma analizza le immagini microarray ed effettua una quantificazione del livello di espressione del gene; costruisce una griglia per identificare gli spots calcola il valore del gene come un integrale del valore dei pixel non saturati, anche se si possono calcolare come il valore medio del valore dei pixel. Gli spot possono essere selezionati anche manualmente nei casi più critici.

## Screenshot:


TIGR Spotfinder 2.0.4
Ready
X : 681
Y : 329
pixel value in A : 46051
pixel value in B : 35597
Column :
Row : 0

## 3.8 Spot 2.0

**Autori**

Ajay N. Jain, Taku Tokuyasu.

**Organizzazione**

Jain Lab, UCSF; San Francisco, California, USA.

**Sistema operativo e ambiente operativo**

Windows 95/98/NT

**Versione**

2.0

**Licenza**

Gratuita per uso accademico.

**Note**

Il software è fornito in formato eseguibile;è fornito il manuale di utilizzo e dati di prova.

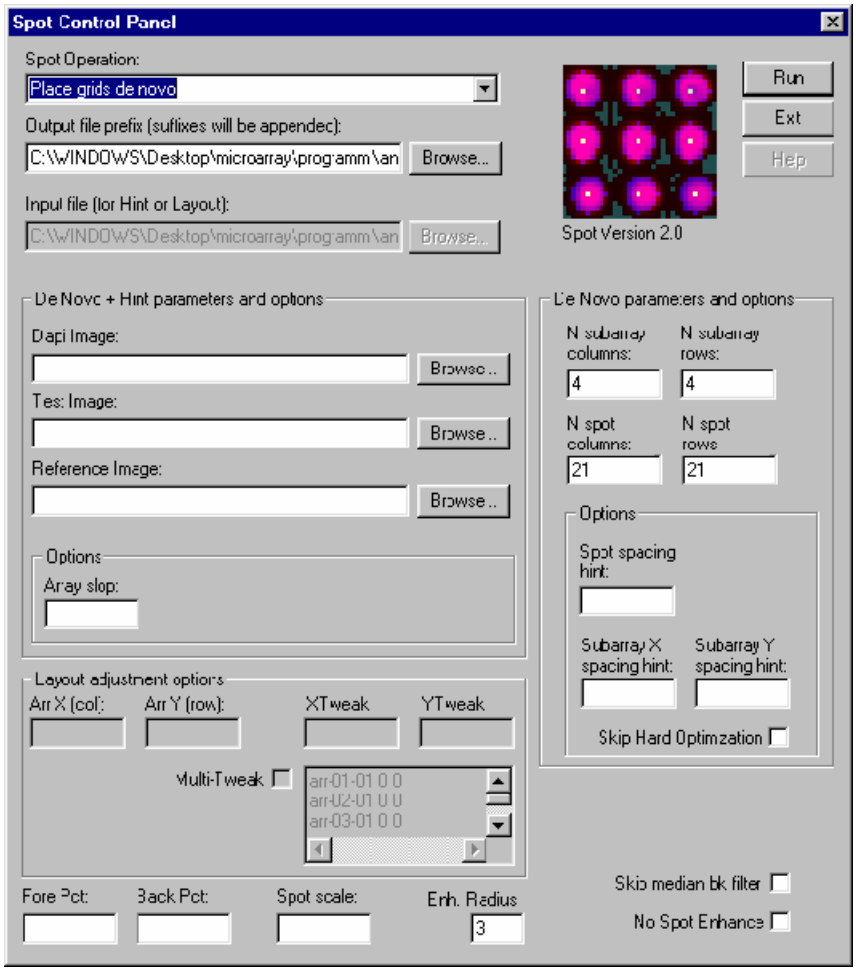
**Sito web:**

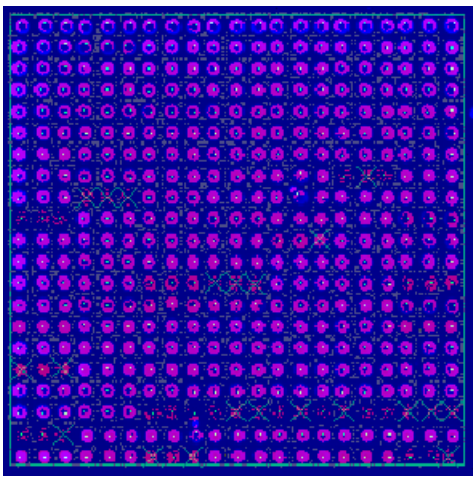
http://jainlab.ucsf.edu/Projects.html

**Funzionamento**

Il programma analizza le immagini microarray ed effettua una quantificazione del livello di espressione del gene. Fornisce un file di testo con i dati statistici per ogni spot.

## Screenshot:


Spot Control Panel
Spot Operation:
Place grids de novo
Output file prefix (suffixes will be appended):
C:\WINDOWS\Desktop\microarray\programm\an
Browse...
Input file (for Hint or Layout):
C:\WINDOWS\Desktop\microarray\programm\an
Browse...
Run
Ext
Help
Spot Version 2.0
De Novo + Hint parameters and options
Dapi Image:
Browse..
Tes: Image:
Browse..
Reference Image:
Browse..
Options
Array slop:
De Novo parameters and options
N subarray columns:
4
N subarray rows:
4
N spot columns:
21
N spot rows
21
Options
Spot spacing hint:
Subarray X spacing hint:
Subarray Y spacing hint:
Skip Hard Optimzation
Layout adjustment options
Arr X (col):
Arr Y (row):
XTweak
YTweak
Multi-Tweak
arr-01-01 0 0
arr-02-01 0 0
arr-03-01 0 0
Fore Pct:
Back Pct:
Spot scale:
Enh. Radius
3
Skip median bk filter
No Spot Enhance



# CAPITOLO IV

# Software per la normalizzazione ed il data mining

## 4.1 Introduzione

Analizzato un singolo microarray relativo ad un particolare tipo di cellula il passo successivo nella ricerca è quello di confrontare più microarray relativi allo stesso tipo di cellula ma in condizioni di sperimentazioni diverse. I dati complessivi vengono memorizzati in una matrice in cui le colonne fanno riferimento agli esperimenti mentre le righe si riferiscono ai geni di quel particolare tipo di cellula.

La maggioranza dei programmi utilizza lo stesso formato per la matrice in ingrasso che possiamo vedere nelle immagini seguenti:

| YORF | 0 minutes | 30 minutes | 1 hour | 2 hours | 4 hours |
|---|---|---|---|---|---|
| YAL001C | 1 | 1.3 | 2.4 | 5.8 | 2.4 |
| YAL002W | 0.9 | 0.8 | 0.7 | 0.5 | 0.2 |
| YAL003W | 0.8 | 2.1 | 4.2 | 10.1 | 10.1 |
| YAL005C | 1.1 | 1.3 | 0.8 | | 0.4 |
| YAL010C | 1.2 | 1 | 1.1 | 4.5 | 8.3 |

oppure

| YORF | NAME | GWEIGHT | GORDER | 0 | 30 | 1 | 2 | 4 |
|---|---|---|---|---|---|---|---|---|
| EWEIGHT | | | | 1 | 1 | 1 | 1 | 0 |
| EORDER | | | | 5 | 3 | 2 | 1 | 1 |
| YAL001C | TFIIC 138 KD SUBUNIT | 1 | 1 | 1 | 1.3 | 2.4 | 5.8 | 2.4 |
| YAL002W | UNKNOWN | 0.4 | 3 | 0.9 | 0.8 | 0.7 | 0.5 | 0.2 |
| YAL003W | ELONGATION FACTOR EF1-BETA | 0.4 | 2 | 0.8 | 2.1 | 4.2 | 10.1 | 10.1 |
| YAL005C | CYTOSOLIC HSP70 | 0.4 | 5 | 1.1 | 1.3 | 0.8 | | 0.4 |

Il secondo formato contiene anche le informazioni sul nome del gene.

Queste matrici in ingresso ai programmi vengono visualizzate come immagini associando ad ogni valore dell'elemento della matrice un colore con una legge di corrispondenza che permetta di evidenziare meglio i dati.

In output questi programmi forniscono immagini o file in formato testo in cui sono indicate le relazioni tra i geni o tra gli esperimenti individuate.

## 4.2 Cluster and TreeView

**Autori**

Michael Eisen's lab;

**Organizzazione**

Lawrence Berkeley National Lab (LBNL), University of California at Berkeley(UCB), Berkeley ,California – USA.

**Sistema operativo e ambiente operativo**

Windows 95/98/NT

**Versione**

2.20

**Licenza**

Gratuita per uso accademico.

**Note**

Il software è fornito in formato eseguibile con file di installazione, è fornito il manuale utente, dati di esempio e codice sorgente in linguaggio C++. Per la visualizzazione dei dati in forma grafica viene utilizzato il programma TreeView dello stesso autore descritto in seguito.

**Sito web:**

http://rana.lbl.gov/EisenSoftware.htm

**Funzionamento**

Il programma accetta dati in ingresso nel formato specificato nell'introduzione del capitolo.

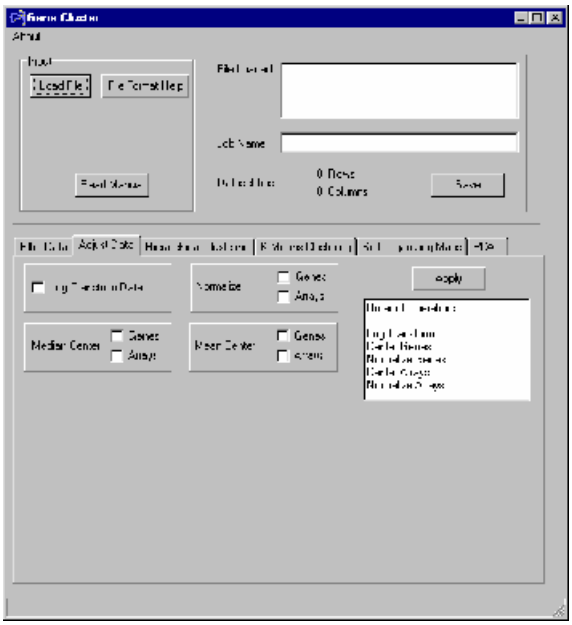
In output il programma fornisce dei file di testo con i risultati dell'analisi che per essere esaminati graficamente sono posti in input al programma TreeView.

I dati prima di essere sottoposti ad analisi di tipo data mining possono subire altri tipi di trasformazioni come la normalizzazione o il filtraggio.

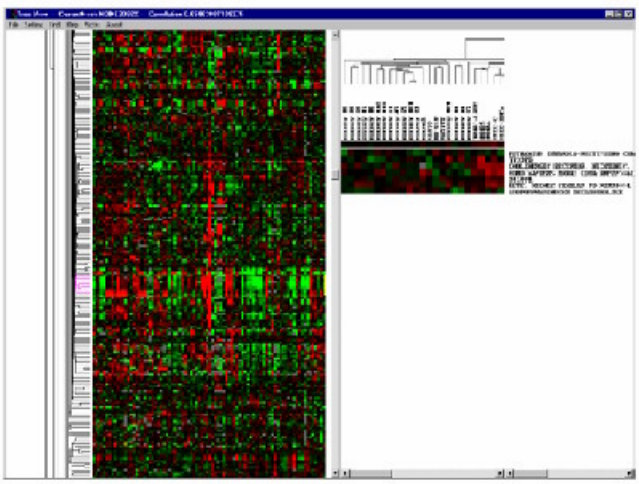
Le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis, la creazioni di mappe autoorganizzanti ( SOM ) e la principal componont analisis (PCA) oltre che varie normalizzazioni e filtraggi.

**Screenshots:**

Cluster:



TreeView:



## 4.3 Genesis

**Autore**

Alexander Sturn;

**Organizzazione**

Bioinformatics Group, Institute of Biomedical Engineering, Graz University of Technology; Graz, Austria.

**Sistema operativo e ambiente operativo**

Java 1.4.2 e Java3D 1.3.1 su qualsiasi sistema operativo che supporta Java.

**Versione**

1.2.2

**Licenza**

Gratuito per uso accademico.

**Note**

Il software è fornito in formato eseguibile con file di installazione, è fornito il manuale utente e dati di esempio. Ha bisogno dell'installazione del pacchetto Java3D 1.3.1

**Sito web:**

http://genome.tugraz.at/Software/GenesisCenter.html

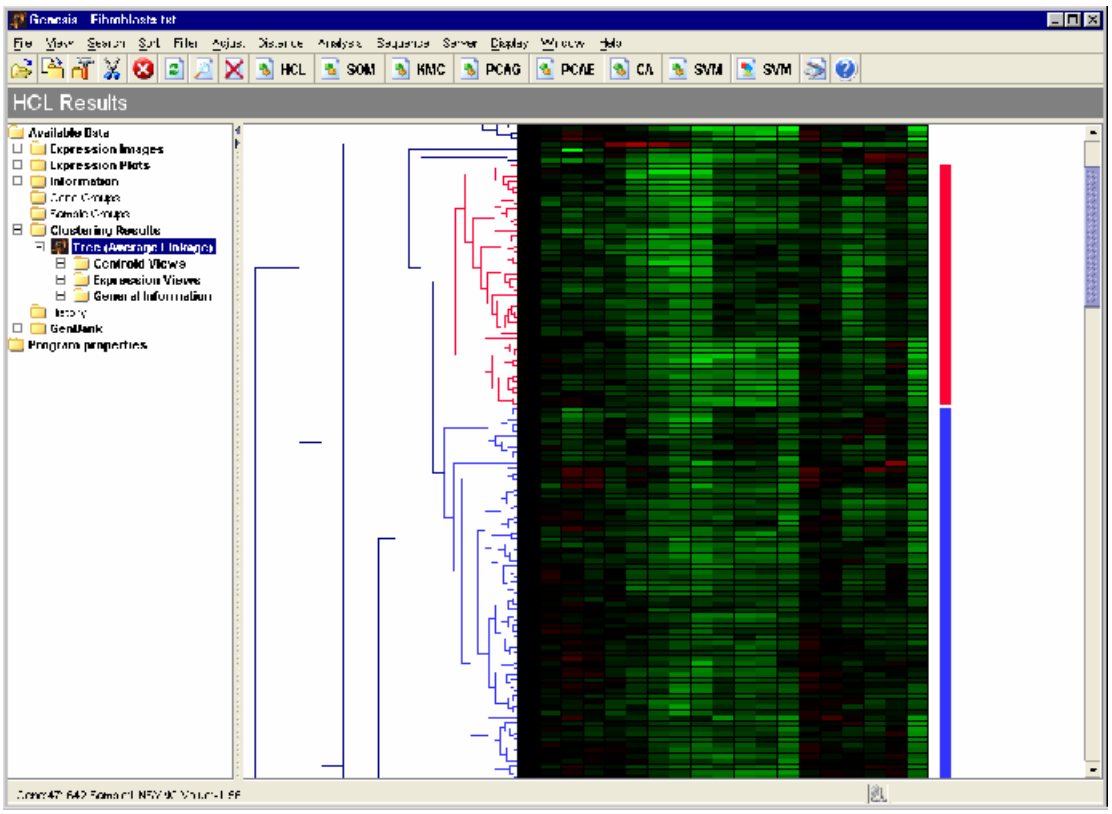
**Funzionamento**

Il programma accetta dati in ingresso nel formato specificato nell'introduzione del capitolo.

I dati prima di essere sottoposti ad analisi di tipo data mining possono subire altri tipi di trasformazioni come la normalizzazione o il filtraggio.

Le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis, la creazioni di mappe autoorganizzanti ( SOM ) e la principal componont analisis (PCA) ed anche la classificazione supervisionata con la support vector machine (SVM).

**Screenshot:**



## 4.4 J-express

**Autori**

Bjarte Dysvik e Inge Jonassen;

**Organizzazione**

Dept. of Informatics, University of Bergen; Bergen, Norvegia.

**Sistema operativo e ambiente operativo**

Java testato su varie piattaforme Windows e Unix

**Versione**

1.1

**Licenza**

Gratuito per uso accademico.

**Note**

Il software è fornito in formato eseguibile con file di installazione, è fornito il manuale utente e dati di esempio.

**Sito web:**

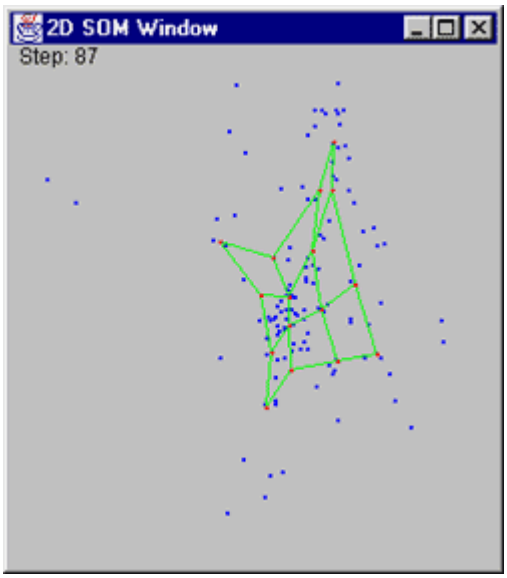
http://www.ii.uib.no/~bjarted/jexpress/

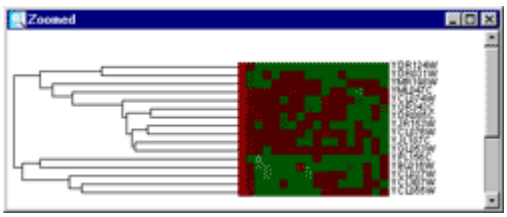
**Funzionamento**

Il programma accetta dati in ingresso nel formato specificato nell'introduzione del capitolo.

Le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis, la creazioni di mappe autoorganizzanti ( SOM ) e la principal componont analisis (PCA).

**Screenshots:**

## 4.5 MAExplorer - MicroArray Explorer

**Autore**

Peter F. Lemkin;

**Organizzazione**

Laboratory of Experimental and Computational Biology (LECB), Center for Cancer Research, National Cancer Institute(NCI) – Frederick, Maryland – USA.

**Sistema operativo e ambiente operativo**

Java testato su varie piattaforme Windows e Unix

**Versione**

0.96.33.06

**Licenza**

Gratuita: Mozilla Public License 1.1

**Note**

Il software è fornito in formato eseguibile con file di installazione, è fornito il manuale utente e dati di esempio. E' fornito un programma di supporto, Cvt2Mae , per la conversione dei dati in ingresso dai vari formati testuali a quello utilizzato dal programma.

**Sito web:**

http://maexplorer.sourgeforge.net/

oppure

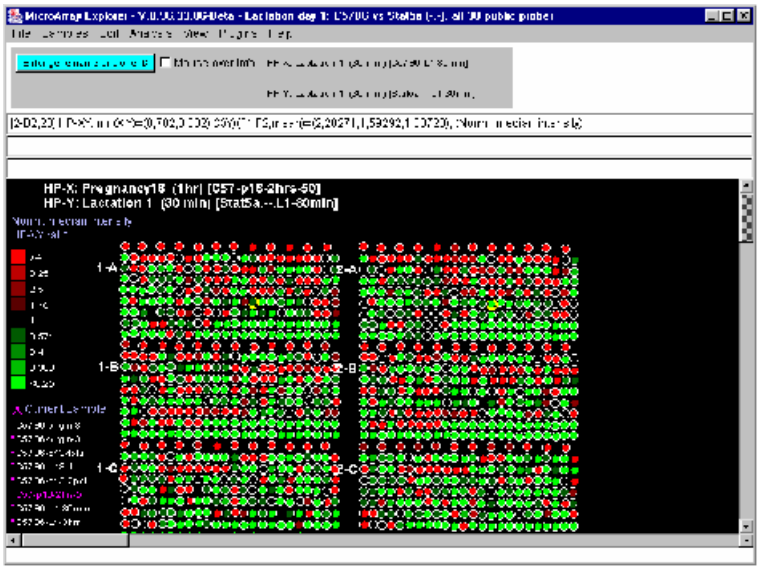
http://www.lecb.ncifcrf.gov/MAExplorer

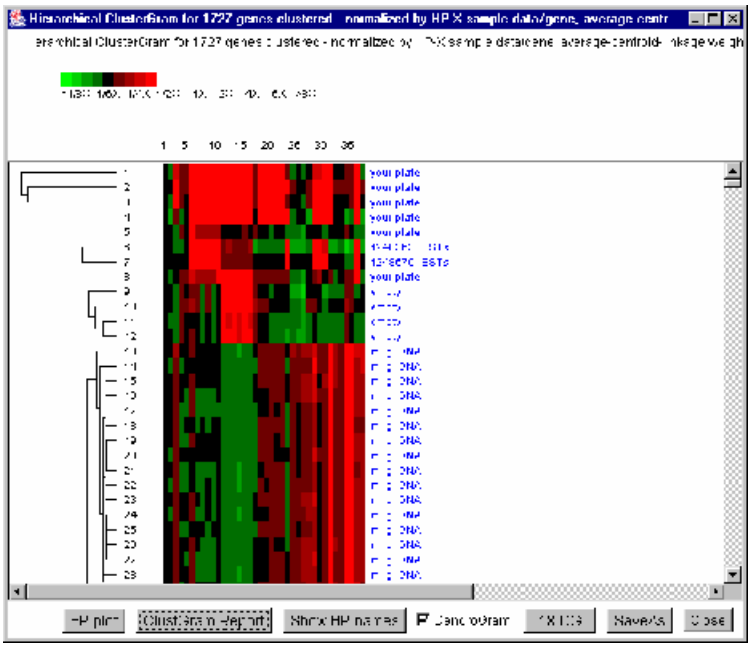
**Funzionamento**

Il programma mette a disposizione metodi per la normalizzazione e il filtraggio dei dati e la visualizzazione grafica mediante istogrammi.

Le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis e la K-median analisis.

**Screenshots:**





## 4.6 TIGR Multiple Experiment Viewer (MEV)

**Autori**

Alexander I. Saeed, Nirmal Bhagabati, John Braisted, Alexander Sturn e John Quackenbush;

**Organizzazione**

The Institute of Genomic Research (TIGR); Rockville, Maryland - USA

**Sistema operativo e ambiente operativo**

Java 1.4.1 testato su varie piattaforme Windows e Unix , Java 3D 1.3.1

**Versione**

2.2

**Licenza**

Gratuita (Open Source)

**Note**

Il software è fornito in formato eseguibile con file di installazione, è fornito il manuale utente e dati di esempio.

Il file di avvio TMEV.BAT va modificato nel seguente modo:

questa parte:

```
set ClassPath = TMEV.jar; Lib/algorithms.jar; Lib/algorithms-gui.jar ;Lib/algorithms-impl.jar; Lib/algorithms-gui-impl.jar; Lib/Images.jar; Lib/jaxp.jar; Lib/parser.jar; Lib/jconn2.jar; Lib/jai_core.jar; Lib/jai_codec.jar; Lib/xerces.jar; Lib/base64.jar; Lib/HTTPClient.jar; Lib/JSciPartial.jar; Lib/JSciCore.jar; Lib/jama.jar; Lib/dialogHelp.jar; Lib/normalization.jar;
java -Xmx256m -cp %ClassPath% org.tigr.microarray.mev.TMEV
```

va sostituita con questa:

java -Xmx256m -cp TMEV.jar;Lib/algorithms.jar;Lib/algorithms-gui.jar;Lib/algorithms-impl.jar;Lib/algorithms-gui-impl.jar;Lib/Images.jar;Lib/jaxp.jar;Lib/parser.jar;Lib/jconn2.jar;Lib/jai_core.jar;Lib/jai_codec.jar;Lib/xerces.jar;Lib/base64.jar;Lib/HTTPClient.jar;Lib/JSciPartial.jar;Lib/JSciCore.jar;Lib/jama.jar;Lib/dialogHelp.jar;Lib/normalization.jar; org.tigr.microarray.mev.TMEV

**Sito web:**

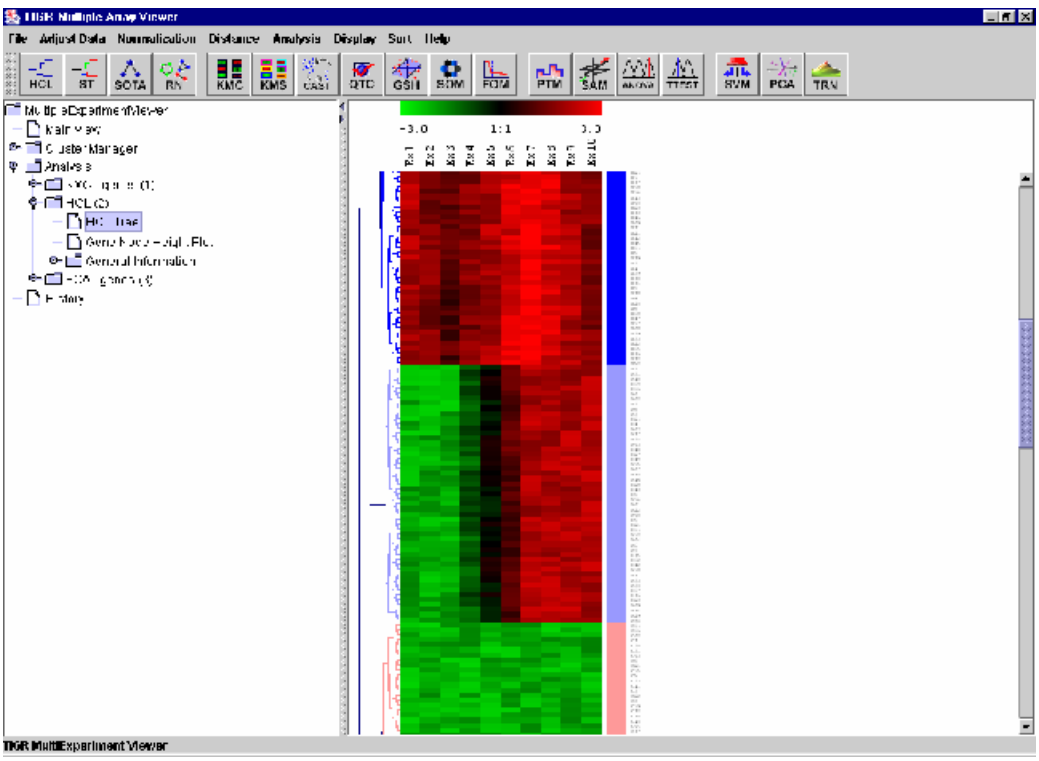
http://www.tigr.org/software/tm4/mev.html

**Funzionamento**

Questa applicazione Java è progettata per permettere l'analisi di dati microarray per identificare i modelli di espressione dei geni e la loro differenziazione.

Il programma mette a disposizione metodi per la normalizzazione e il filtraggio dei dati, le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis, mappe autoorganizzanti (SOM), principal component analisis (PCA) e molte altre.

**Screenshot:**



## 4.7 AMIADA (Analysis of Microarray Data)

**Autori**

Xuhua Xia;

**Organizzazione**

Department of Biology(http://www.bio.uottawa.ca/eng/welcome.php ), University of Ottawa; Ontario, Canada.

**Sistema operativo e ambiente operativo**

Windows 98/NT/2000

**Versione**

2.0.7

**Licenza**

Gratuita per uso accademico

**Note**

Il software è fornito in formato eseguibile, è disponibile anche il manuale utente.

**Sito web:**

http://aix1.uottawa.ca/%7Exxia/software/amiada.htm

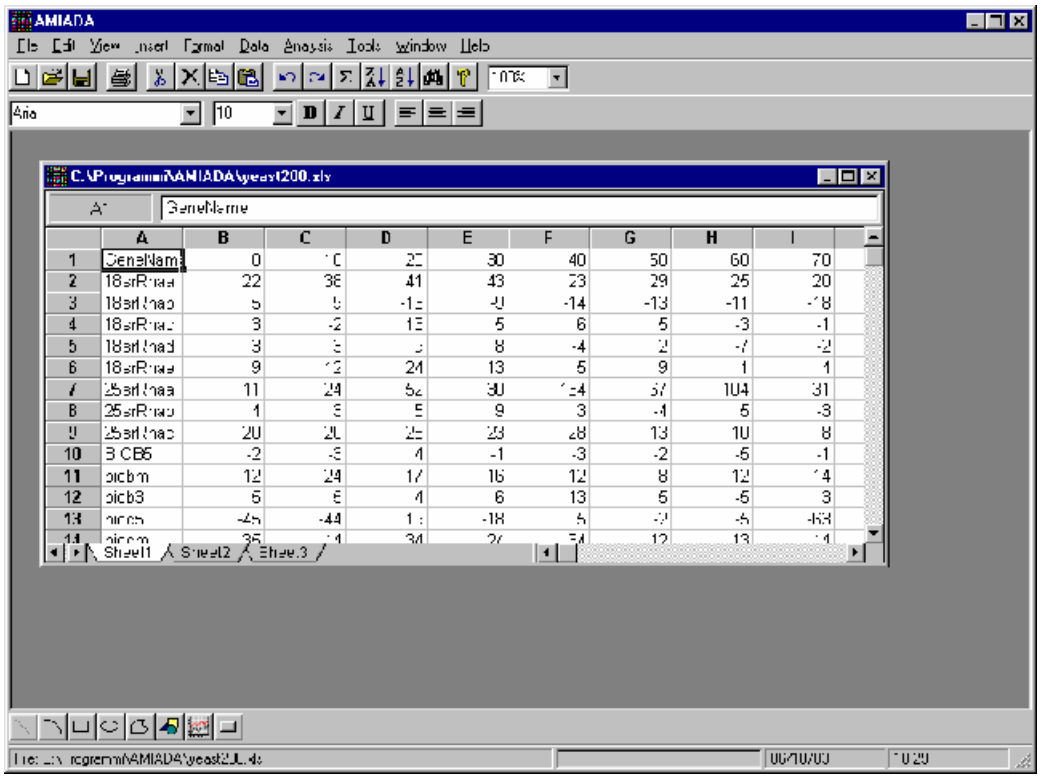
**Funzionamento**

I dati in ingresso devono essere in formato EXCEL; tramite questo programma però si possono convertire gli altri formati in questo.

Il programma permette l'organizzazione, l'esplorazione, la visualizzazione e l'analisi dei dati degli esperimenti microarray; Utilizza una interfaccia utente

EXCEL-like ed è in grado di effettuare cluster analysis, PCA, e varie normalizzazioni.

**Screenshot:**



## 4.8 R-maanova

**Autori**

Hao Wu , Gary A. Churchill;

**Organizzazione**

Churchill Statistical Genetics Group, The Jackson Laboratory; Bar Harbor, Maine – USA.

**Sistema operativo e ambiente operativo**

Ambiente Matlab, oppure R, su sistemi operativi Windows e Linux

**Versione**

Matlab : 2.0

R : 0.91

**Licenza**

Gratuita.

**Note**

Il software è implementato come una libreria di funzioni che possono essere utilizzate all'interno degli ambienti Matlab ed R per l'analisi dei microarray. Sono fornite le informazioni di installazione e vari esempi sull'utilizzo delle funzioni. Delle funzioni è fornito il codice sorgente.

**Sito web:**

http://www.jax.org/research/churchill/software/anova/index.html

**Funzionamento**

Le funzioni permettono la visualizzazione dei dati, la verifica della qualità , la trasformazione degli stessi, la ANalysis Of Variance, e la cluster analisis.

## 4.9 Genecluster

**Autori**

Keith Ohm, Michael Angelo;

**Organizzazione**

Whitehead Institute Centre for genome research; Cambridge, Massachusetts, USA

**Sistema operativo e ambiente operativo**

Tutti i sistemi operativi che supportano Java, JRE 1.3.1 o superiori

**Versione**

2.0

**Licenza**

Gratuita  per uso accademico.

**Note**

Il programma è fornito in formato eseguibile e sono fornite le istruzioni di installazione e funzionamento.

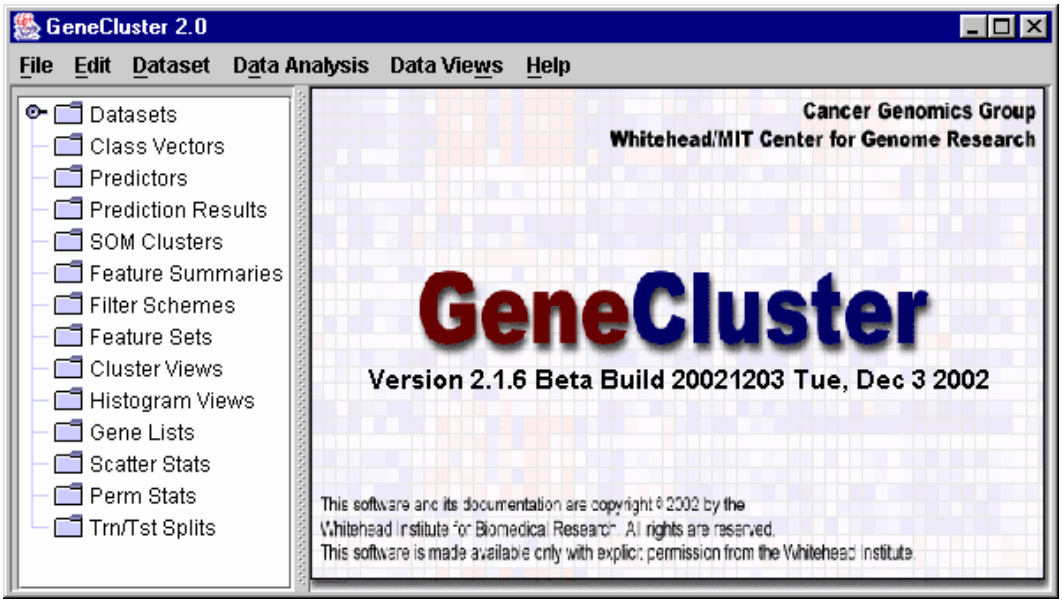
**Sito web:**

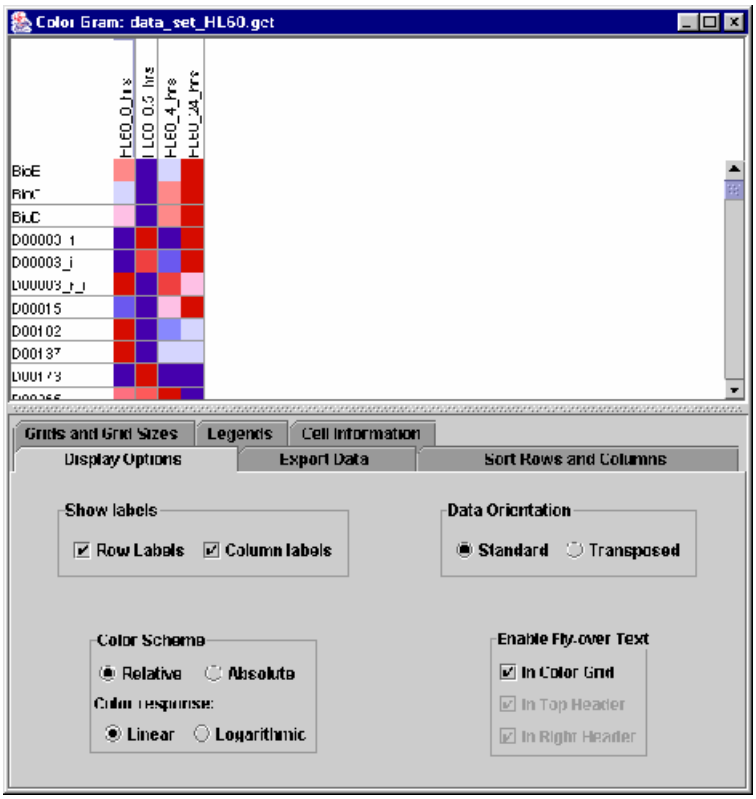
http://www-genome.wi.mit.edu/cancer/software/software.html

**Funzionamento**

Il programma mette a disposizione metodi per la normalizzazione e il filtraggio dei dati, le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, SOM ed altre analisi supervisionate e non.

## Screenshots:



## 4.10 Clustfavor

**Autore**

Leif E. Peterson;

**Organizzazione**

Baylor College of Medicine; Houston, Texas, USA.

**Sistema operativo e ambiente operativo**

Windows 95/98/NT/2000/XP

**Versione**

6.07

**Licenza**

Gratuita per uso accademico.

**Note**

Il programma è fornito in formato eseguibile e sono fornite le istruzioni di installazione e funzionamento.

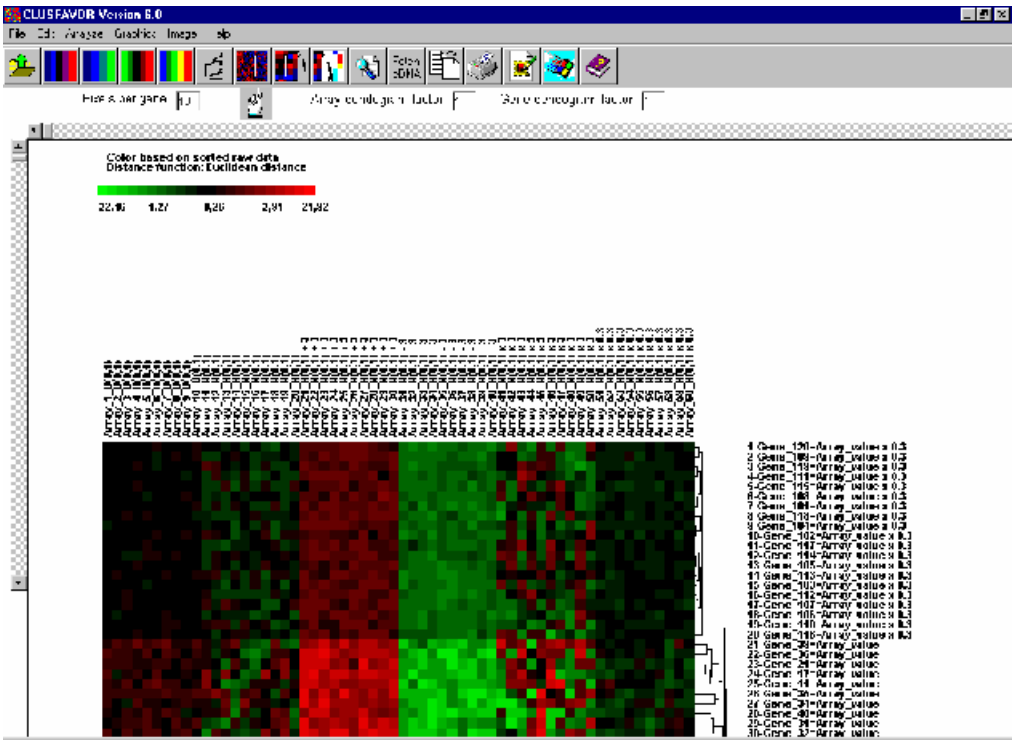
**Sito web:**

http://mbcr.bcm.tmc.edu/genepi/

**Funzionamento**

Il programma mette a disposizione metodi per la normalizzazione e il filtraggio dei dati, le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, principal component analisis (PCA) ed altre.

## Screenshot:


CLUSFAVOR Version 6.0
File Edit Analyze Graphics Images Help
Color based on sorted raw data
Distance function: Euclidean distance

# CAPITOLO V

# Software per la visualizzazione delle reti di interazione tra geni

## 5.1 Introduzione

Questi programmi utilizzano non solo i dati forniti dagli esperimenti microarray ma anche conoscenze sul funzionamento della cellula già acquisite, in modo da analizzare e visualizzare le reti di interazione tra geni gia conosciute e permettono la creazione di nuove reti da sottoporre a successivi studi.

## 5.2 Cytoscape

**Autori**

Tray Ideker, Benno Schwikowski, Chris Sander;

**Organizzazione**

The Institute for Systems Biology, The University of California at San Diego, The Memorial Sloan-Kettering Cancer Center; San Diego, California – USA.

**Sistema operativo e ambiente operativo**

Java 1.4.1 testato su varie piattaforme Windows e Unix

**Versione**

1.1

**Licenza**

Gratuita (Open Source)

**Note**

Il software è fornito in formato eseguibile con file di installazione; è fornito il manuale utente, dati di esempio e il codice sorgente in Java.

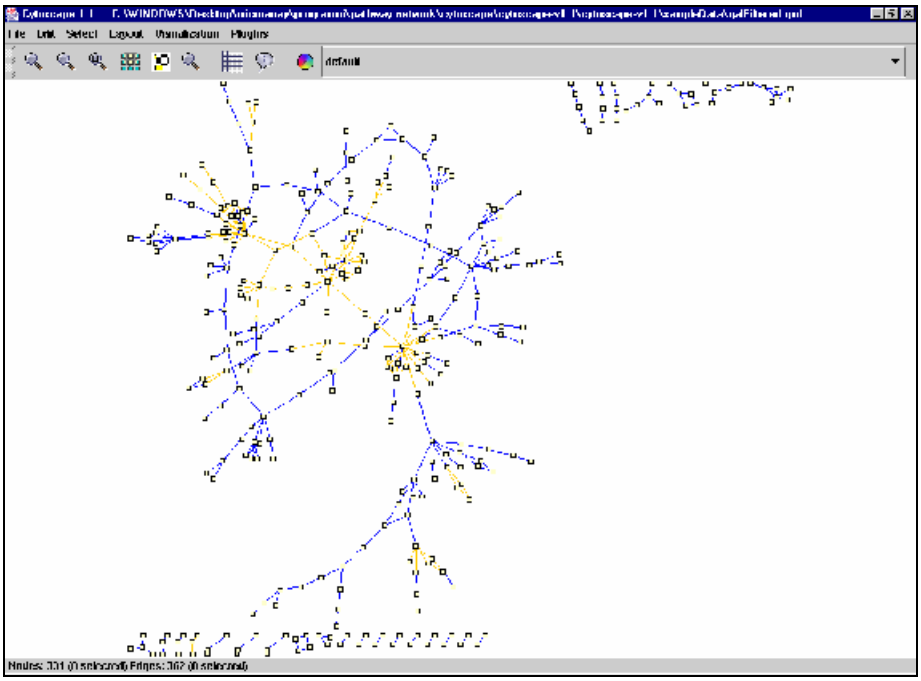
**Sito web:**

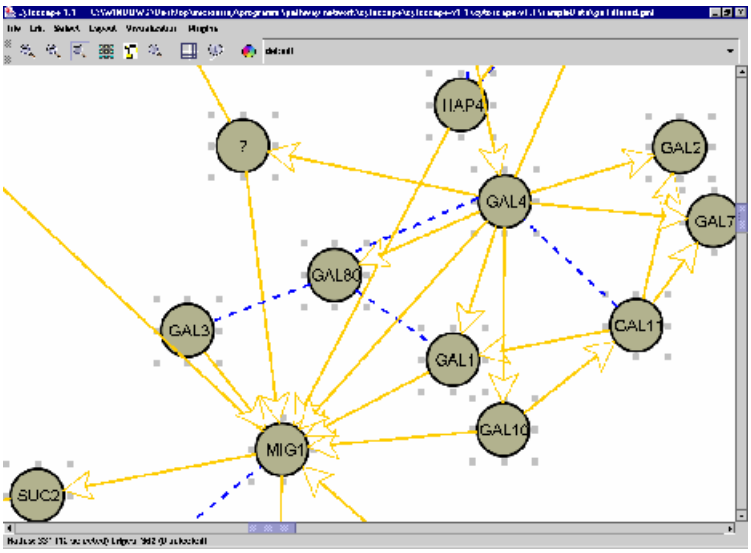
http://www.cytoscape.org/

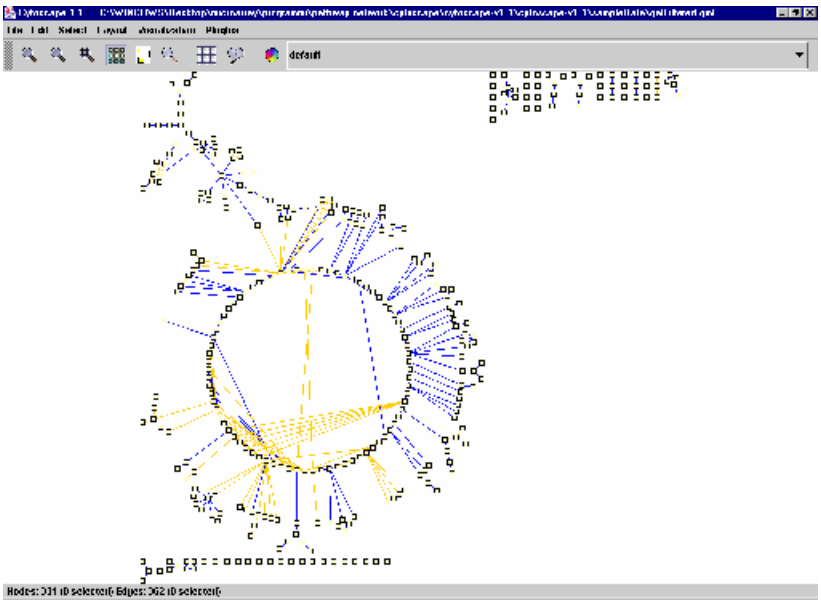
**Funzionamento**

Il programma visualizza le reti delle interazioni molecolari e integra questi dati con i dati sull'andamento delle espressioni dei geni fornito dai microarray. Il programma fornisce un algoritmo che cerca di individuare la particolare rete che controlla i cambiamenti delle espressioni dei geni osservate con i microarray. La visualizzazione delle reti è fatta con un grafo, con geni, proteine e molecole rappresentate come nodi e le interazioni tra questi rappresentati dai collegamenti tra i nodi.

**Screenshots:**




default
IIAP4
?
GAL2
GAL4
GAL7
GAL80
GAL3
CAL11
GAL1
GAL10
MIG1
SUC2


default

## 5.3 GenMAPP (Gene MicroArray Pathway Profiler)

**Autori**

Steven C. Lawlor, Bruce Conklin, Kam Dahlquist

**Organizzazione**

Gladstone Institutes, University of California at San Francisco; San Francisco, California – USA.

**Sistema operativo e ambiente operativo**

Windows 98/2000/NT

**Versione**

1.0

**Licenza**

Gratuita

**Note**

Il software è fornito in formato eseguibile con file di installazione, è fornito il manuale utente e dati di esempio.

**Sito web:**

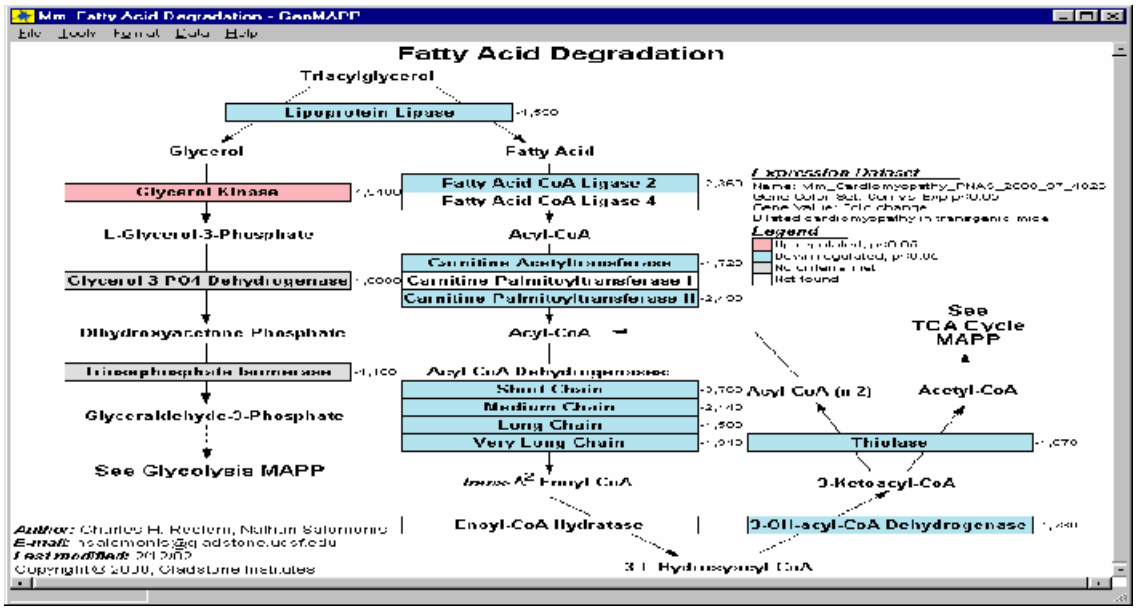
http://www.genmapp.org/

**Funzionamento**

GenMAPP è un'applicazione progettata per la visualizzazione delle espressioni di geni all'interno di mappe rappresentanti le funzioni biologiche della cellula. Il programma permette la visualizzazione delle espressioni dei geni in un contesto biologico attraverso un formato grafico ed intuitivo. Le mappe mostrano le relazioni tra i geni e i prodotti che da essi derivano. Queste mappe,

rappresentano funzioni biologiche standard, rese pubbliche da varie organizzazioni come il Gene Ontology Progect.

**Screenshot:**



## 5.4 Osprey Network Visualization System

**Autori**

Bobby-Joe Breitkeutz, Chris Stark, Mike Tyers;

**Organizzazione**

Mount Sinai Hospital; Toronto, Canada.

**Sistema operativo e ambiente operativo**

Windows 98/NT/2000, Unix/Linux , Mac OS

**Versione**

1.0

**Licenza**

Gratuita

**Note**

Il software è fornito in formato eseguibile con file di installazione, è fornito il manuale utente e dati di esempio.

**Sito web:**

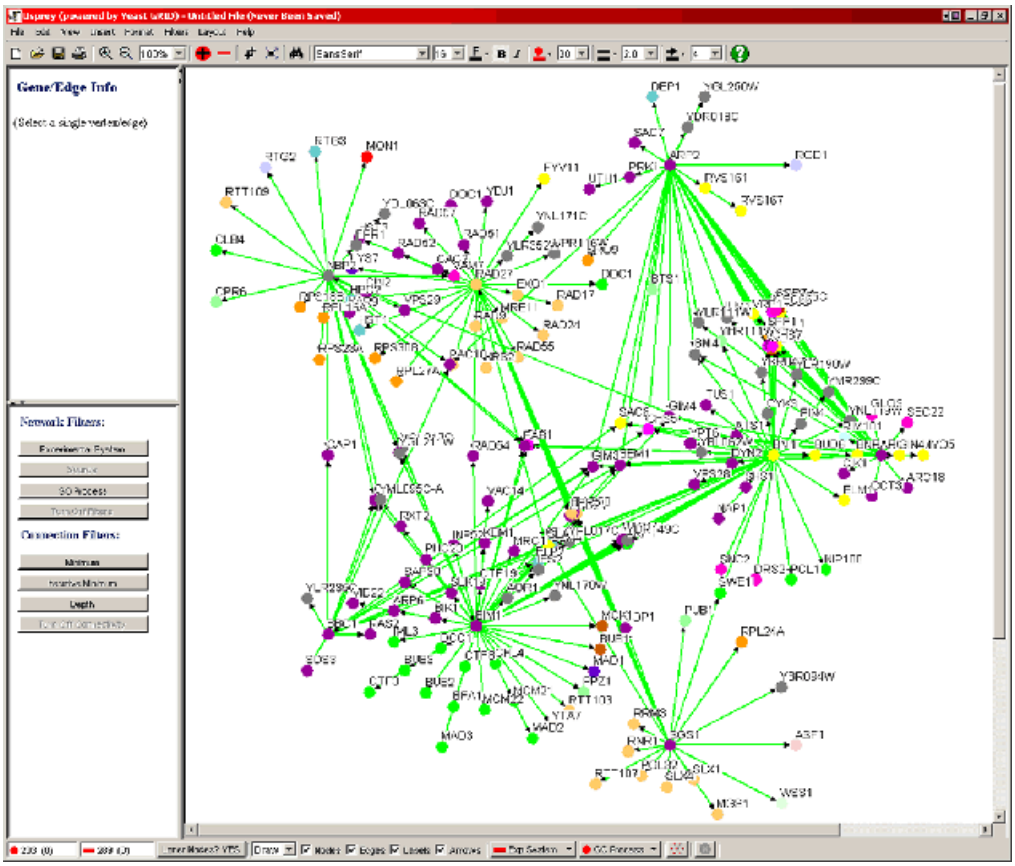
http://biodata.mshri.on.ca/osprey

**Funzionamento**

Il programma fornisce gli strumenti per la visualizzazione e la manipolazione di complesse reti di interazioni tra i geni. Il codice dei colori dei nodi rappresenta la funzionalità del gene anche le interazioni hanno codici di colore per individuare il tipo di interazione. Il programma utilizza collegamenti a data base pubblici on-

line per ottenere le informazioni sul tipo di gene e sulle interazioni in cui è coinvolto.

**Screenshot:**



# CAPITOLO VI

# Software web-based

## 6.1 Introduzione

Questa particolare classe di programmi è definita non dalla funzione che realizza ma dal particolare mezzo che utilizza; questi programmi risiedono su server remoti appartenenti alle organizzazioni che li hanno progettati e mettono a disposizione le loro elaborazioni on-line attraverso internet utilizzando un semplice browser.

I dati da analizzati vengono richiamati dal server (upload) che poi restituisce (download) i risultati che possono essere sia file in formato testo sia in formato grafico.

## 6.2 Engene

**Organizzazione**

Computer Architecture Department, Universidad de Malaga; Malaga, Spagna

**Sistema operativo e ambiente operativo**

Qualsiasi sistema che permetta l'accesso ad internet e su cui sia installato un web browser.

**Licenza**

Gratuita per uso accademico.

**Note**

E' necessario registrarsi come utente prima di poter utilizzare il sistema.

**Sito web:**

http://www.engene.cnb.uam.es/

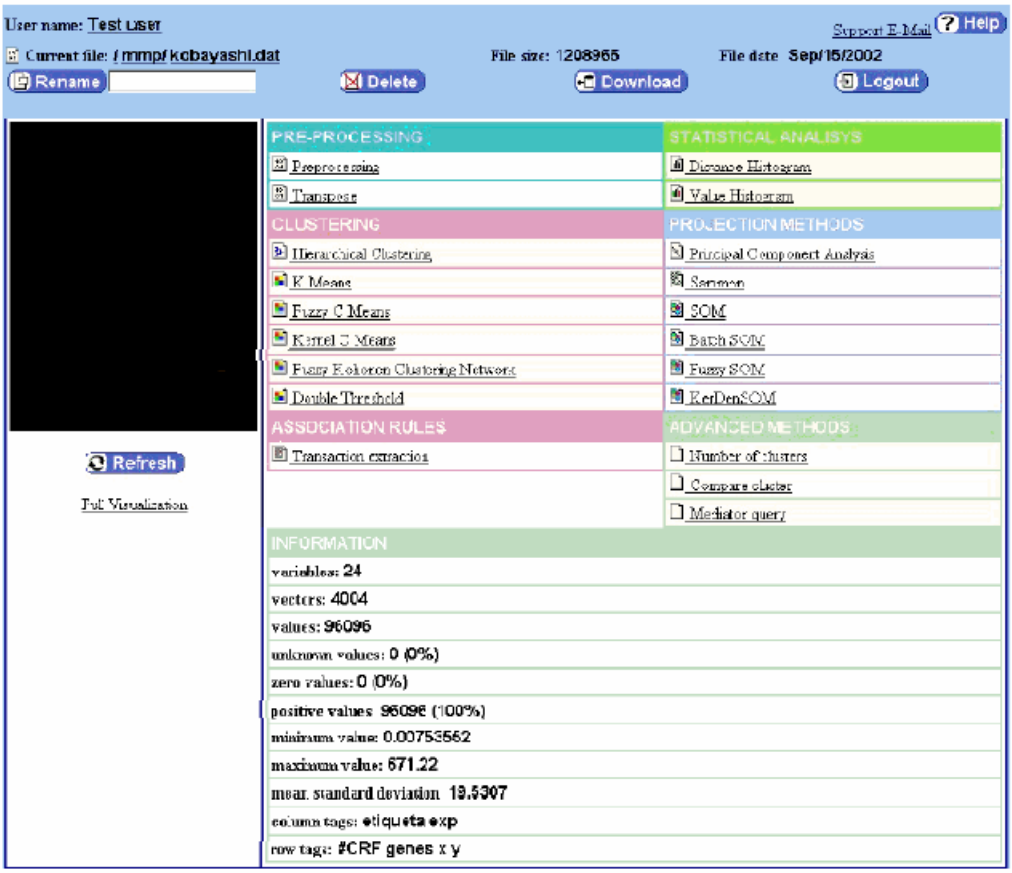
**Funzionamento**

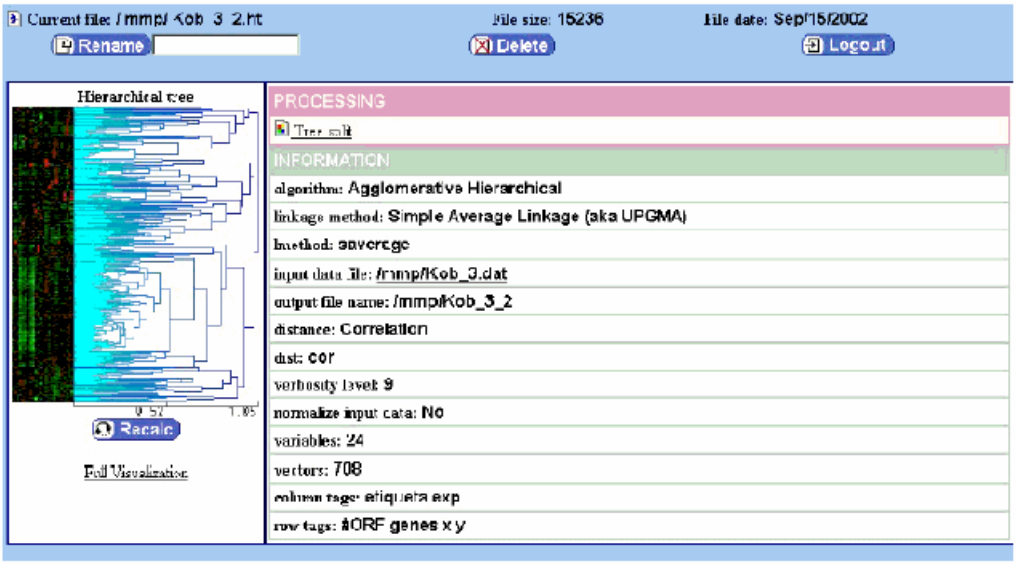
Il sistema richiede l'upload dei dati da analizzare, mette a disposizione gli strumenti di analisi e infine consente il download dei risultati.

I dati prima di essere sottoposti ad analisi di tipo data mining possono subire altri tipi di trasformazioni come la normalizzazione o il filtraggio.

Le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis, la creazioni di mappe autoorganizzanti ( SOM ) ed altre.

**Screenshots:**




Current file: /mmp/Kob_3_2.ht
File size: 15236
File date: Sep/15/2002
Rename
Delete
Logout
Hierarchical tree
0.57
1.85
Recalc
Full Visualization
PROCESSING
Tree cdt
INFORMATION
algorithm: Agglomerative Hierarchical
linkage method: Simple Average Linkage (aka UPGMA)
method: average
input data file: /mmp/Kob_3.dat
output file name: /mmp/Kob_3_2
distance: Correlation
dist: cor
verbosity level: 9
normalize input data: No
variables: 24
vectors: 708
column tags: etiquetas exp
row tags: #ORF genes x y

## 6.3 Expression profiler

**Organizzazione**

European Bioinformatics Institute (EBI); Cambridge, United Kingdom.

**Sistema operativo e ambiente operativo**

Qualsiasi sistema che permetta l'accesso ad internet e su cui sia installato un web browser.

**Licenza**

Gratuito

**Note**

Il sito mette a disposizione diversi tools  non solo per l'analisi dei microarray.

**Sito web:**

http://ep.ebi.ac.uk/

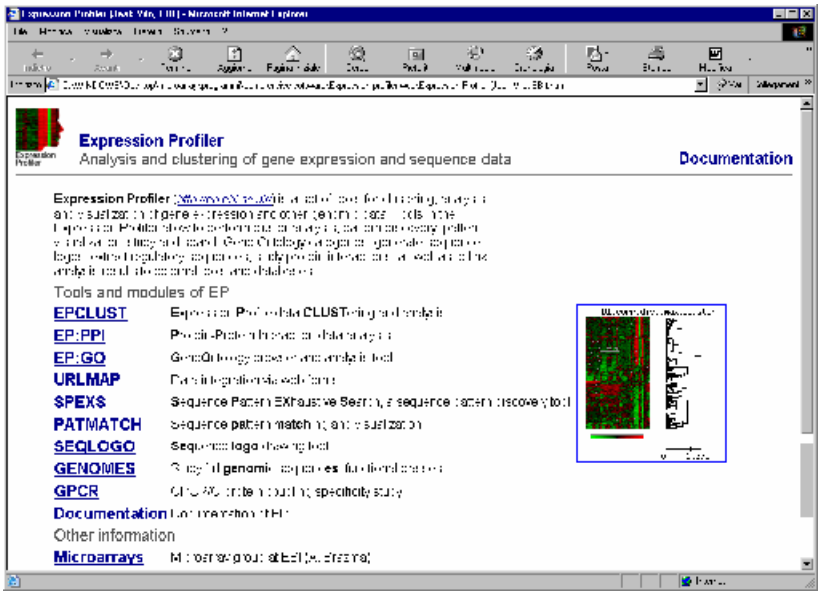
**Funzionamento**

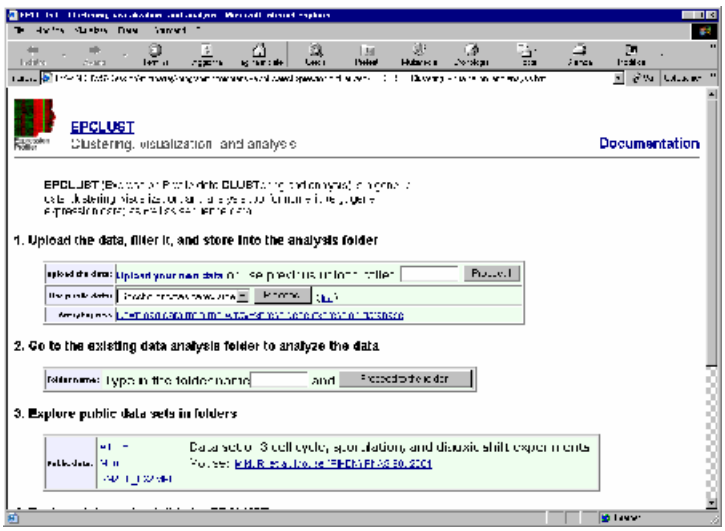
Il sistema richiede l'upload dei dati da analizzare, mette a disposizione gli strumenti di analisi e infine consente il download dei risultati.

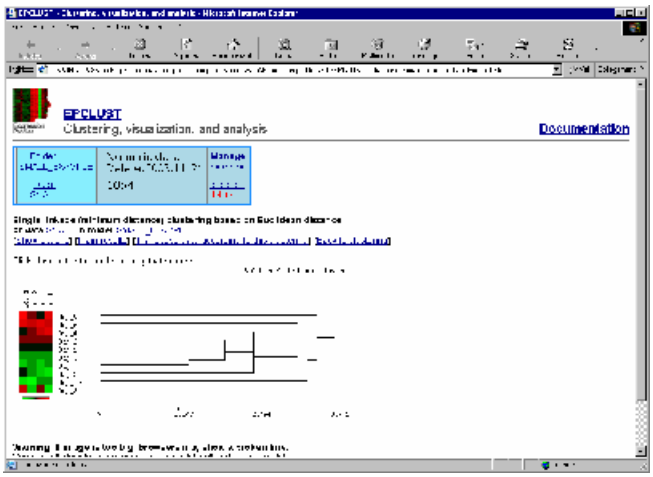
I dati prima di essere sottoposti ad analisi di tipo data mining possono subire altri tipi di trasformazioni come la normalizzazione o il filtraggio.

Le principali analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis ed altre .

## Screenshots:



## 6.4 Gene Expression Data Analysis Tool (GEDA)

**Organizzazione**

UPCI (University of Pittsburgh Cancer Institute), Center for Pathology Informatics ; Pittsburgh - Pennsylvania, USA

**Sistema operativo e ambiente operativo**

Qualsiasi sistema che permetta l'accesso ad internet e su cui sia installato un web browser.

**Licenza**

Gratuita

**Note**

Vengono forniti il manuale di utilizzo e dati di esempio

**Sito web:**

http://bioinformatics.upmc.edu/GE2/GEDA.html

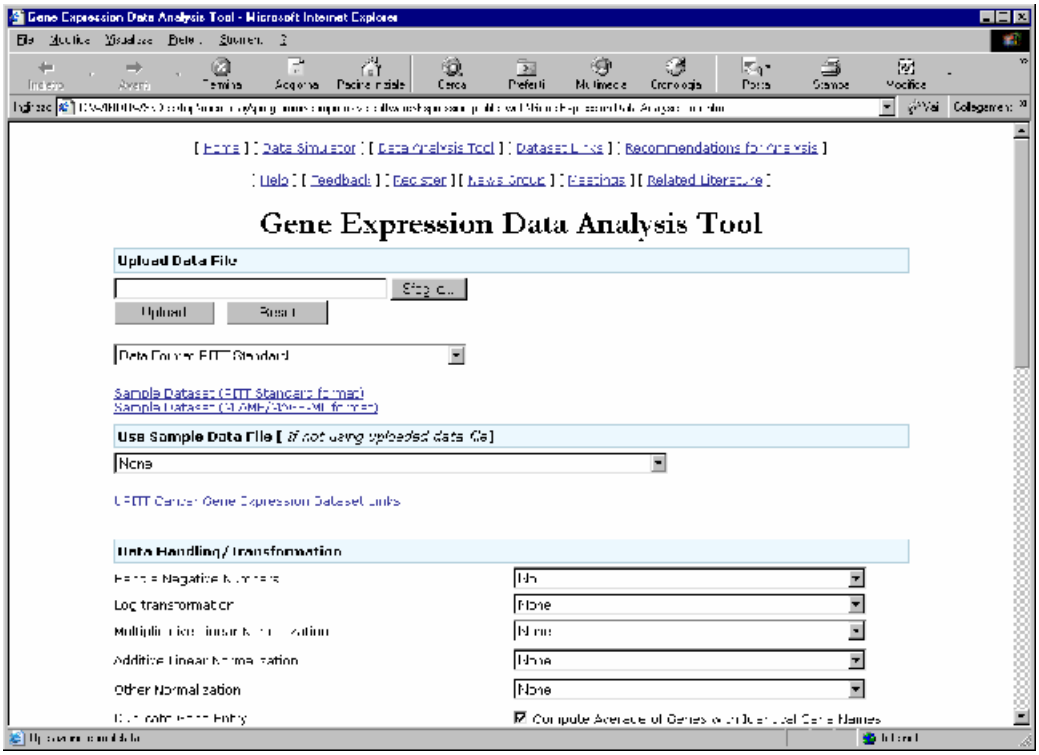
**Funzionamento**

Il sistema richiede l'upload dei dati da analizzare, mette a disposizione gli strumenti di analisi e infine consente il download dei risultati.

I dati prima di essere sottoposti ad analisi di tipo data mining possono subire altri tipi di trasformazioni come la normalizzazione o il filtraggio.

Le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, la K-mean analisis ed altre.

## Screenshot:


Gene Expression Data Analysis Tool - Microsoft Internet Explorer
Home
Data Simulator
Data Analysis Tool
Recommendations for Analysis
Help
Feedback
Register
News Group
Citations
Related Literature
Gene Expression Data Analysis Tool
Upload Data File
Upload
Reset
Sample Dataset (EDIT Standard format)
Use Sample Data File [ If not using uploaded data file]
None
Data Handling / Transformation
Log transformation
None
Other Normalization
None

## 6.5 GEPAS (Gene Expression Pattern Analysis Suite)

**Organizzazione**

Bioinformatics Unit, CNIO (Centro Nacional de Investigaciones Oncologicas); Madrid, Spagna.

**Sistema operativo e ambiente operativo**

Qualsiasi sistema che permetta l'accesso ad internet e su cui sia installato un web browser.

**Licenza**

Gratuita

**Note**

Vengono forniti il manuale di utilizzo e dati di esempio

**Sito web:**

http://gepas.bioinfo.cnio.es/

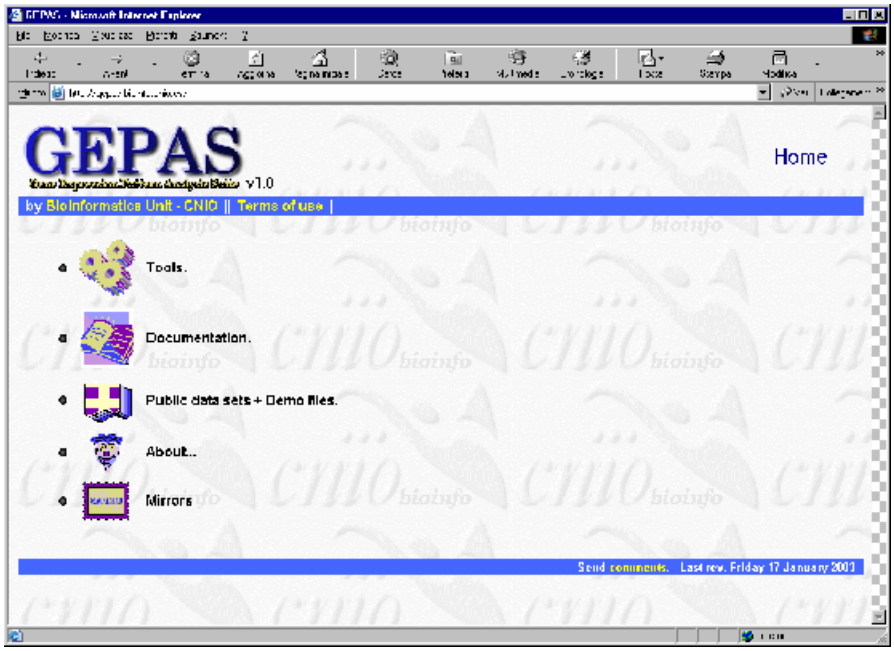
**Funzionamento**

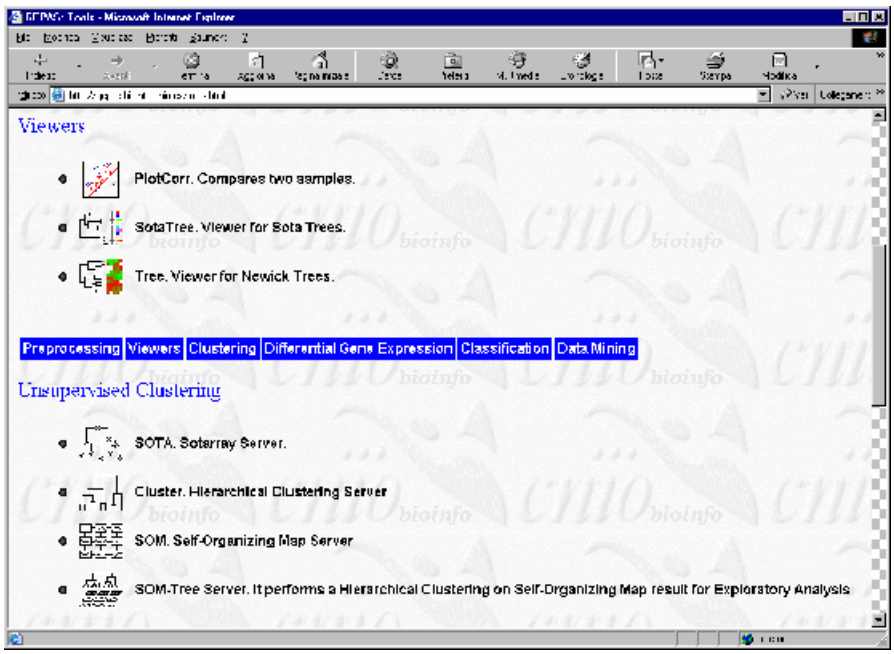
Il sistema richiede l'upload dei dati da analizzare, mette a disposizione gli strumenti di analisi e infine consente il download dei risultati.

I dati prima di essere sottoposti ad analisi di tipo data mining possono subire altri tipi di trasformazioni come la normalizzazione o il filtraggio.

Le analisi che possono essere effettuate con questo programma sono: il clustering gerarchico, SOM, PVM, ed altre .

## Screenshots:



## 6.6 Dynamic signaling map

**Organizzazione**

Hippron Physiomics Inc., Toronto, Canada.

**Sistema operativo e ambiente operativo**

Qualsiasi sistema che permetta l'accesso ad internet e su cui sia installato un web browser.

**Licenza**

Gratuita

**Note**

E' necessario registrarsi come utente prima di poter utilizzare il sistema.

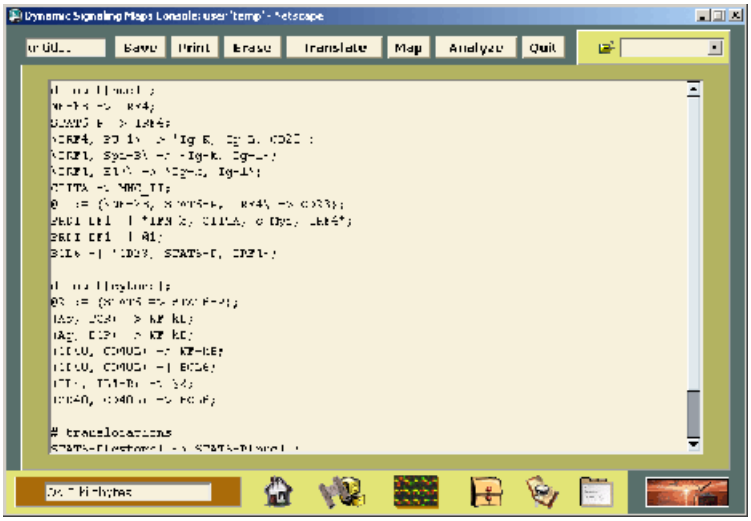
**Sito web:**

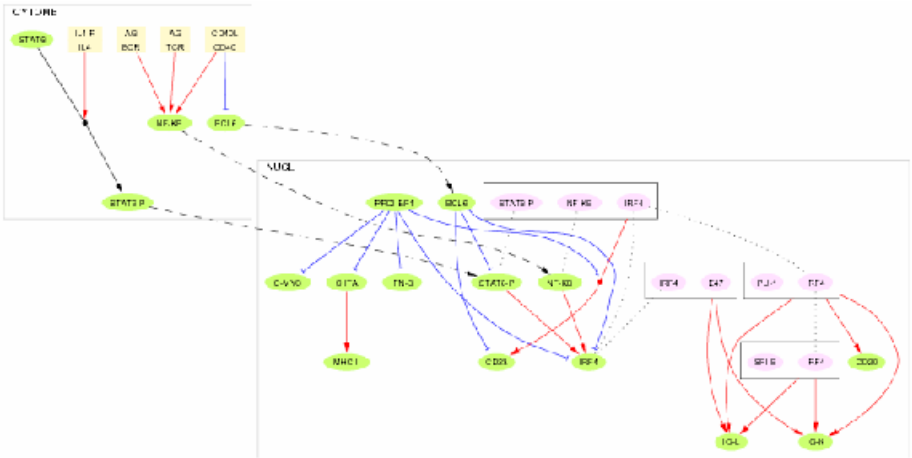
http://www.hippron.com/hippron/index.html

**Funzionamento**

Questa applicazione web-based permette di creare complesse reti di interazioni tra elementi cellulari, queste reti possono essere integrate con i dati su geni e proteine contenuti in data base pubblici. Si possono creare reti di geni dalle matrici di espressione dei geni.

## Screenshots:



# CAPITOLO VII

# Conclusioni

Nella tabella seguente sono elencati tutti i sofware freeware trovati e illustrati nei capitoli precedenti mettendo in evidenza l'ambiente di esecuzione, la disponibilità o meno del codice sorgente e in caso affermativo è indicato il linguaggio con cui è scritto il codice; è mostrata a anche la funzione principale svolta dal codice.

I software sono stati provati su un computer con un AMD 400 Mhz, con 128 MB di memoria RAM e scheda grafica integrata con 8 MB di memoria condivisa.

| **Software** | **Sistema/Ambiente Operativo** | **Codice sorgente** | **Funzione Svolta** |
|---|---|---|---|
| Dapple | Unix/Linux | C++ | Image quantify |
| F-scan | Matlab | Matlab | Image quantify |
| P-scan | Matlab | Matlab | Image quantify |
| GridGrinder | Windows 98/NT/2000 | C++ | Image quantify |
| Spotfinder | Windows 98/NT/2000 | no | Image quantify |
| Spot | Windows 98/NT/2000 | no | Image quantify |
| Scanalyze | Windows 98/NT/2000 | C++ | Image quantify |
| Cluster and Treeview | Windows 98/NT/2000 | C++ | Clustering , normalizzazione |
| Genesis | Java | no | Clustering , normalizzazione |
| J-express | Java | no | Clustering |
| MicroArrayExplorer | Java | no | Clustering , normalizzazione |
| MEV | Java | Java | Clustering , normalizzazione |
| AMIADA | Windows 98/NT/2000 | no | Clustering , normalizzazione |
| Genecluster | Java | no | Clustering , normalizzazione |
| Clustfavor | Windows 98/NT/2000 | no | Clustering , normalizzazione |
| Anova | Matlab / R | Matlab/R | Clustering , normalizzazione |
| Cytoscape | Java | Java | Network Reconstruction |
| Gen Map | Windows 98/NT/2000 | no | Network Reconstruction |
| Osprey | Windows, Unix, Mac | no | Network Reconstruction |

| | | | |
|---|---|---|---|
| | OS | | |
| Engene | WEB | no | Clustering , normalizzazione |
| Expression Profiler | WEB | no | Clustering , normalizzazione |
| GEPAS | WEB | no | Clustering , normalizzazione |
| GEDA | WEB | no | Clustering , normalizzazione |
| Dynamic Signaling Map | WEB | no | Network Reconstruction |

I software elencati mostrano che è possibile approntare un laboratorio a costo zero almeno per quanto riguarda i software che comprenda tutti le categorie di software necessari, dalla analisi delle immagini alla ricostruzione delle reti di interazione dei geni.

Anche se i software funzionano con l'hardware descritto in precedenza è consigliato una maggiore quantità di risorse; un processore più veloce diminuisce i tempi di elaborazione che in alcuni casi hanno raggiunto alcuni minuti; anche la quantità di memoria RAM è da aumentare almeno a 256MB perché la mole di dati da elaborare, specialmente per l'analisi delle immagini, può essere considerevole.

Per le memorie di massa è consigliabile almeno un hard disk 80GB di capacità perché le immagini e le matrici legate ai microarray sono generalmente di grandi dimensioni e quindi per immagazzinare dati relativi a più microarray è necessaria molta memoria.

Per la sezione grafica non sono necessarie grandi risorse per la grafica tridimensionale, più importante sono le risorse per la grafica bidimensionale, quindi è consigliabile una scheda che offra una risoluzione minima di 1024x768 ed un monitor di 17 pollici di diagonale.

Un set completo di programmi uno per ogni categoria è disponibile solo per il sistema operativo Windows; aggiungendo al sistema l'ambiente Java si amplia di molto la scelta dei programmi che si possono utilizzare soprattutto per il data mining.

Bisogna dire che l’ambiente Java è disponibile per molti sistemi operativi quindi anche un sistema operativo Unix/Linux con l’ambiente Java permette l’utilizzo di un completo set di software, uno per ogni categoria.
Sono presenti strumenti di analisi anche per l’ambiente Matlab, che però è a pagamento, ma gli strumenti non coprono tutte le categorie di programmi.
Notiamo che molti programmi forniscono anche il codice sorgente con licenza GPL, questo offre la possibilità di migliorare o ampliare i software che lo permettono e di creare un ambiente integrato per una maggiore collaborazione tra i software.

# Bibliografia


1. GIUDICI P., ”Data Mining”, McGraw-Hill, 2001.
2. BROOKER R. J., ”Genetica. Analisi e Principi”, Zanichelli, 2000.
3. LUSCOMBE N. M., GREENBAUM D., GERSTEIN M., “What is Bioinformatics? A proposed definition and overview of the field“, Method Inform Med 2001; 40:346-358.
4. FARABEE M. J., “On Line biology book”, 2001, (http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html)
5. ”Bioinformatics”, NCBI, 2001, (http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html).
6. BRAZMA A., PARKINSON H., SCHLITT T., SHOJATALAB M., “A quick introduction to elements of biology – cells, molecules, genes, functional genomics, microarrays”, EMBL-EBI, 2001, (http://www.ebi.ac.uk/microarray/biology_intro.html).
7. FRISTENSKY B., “Gene arrays”, 2002, (http://www.umanitoba.ca/faculties/afs/plant_science/courses/39_769/lec12/lec12.1.html).
8. FRISTENSKY B., “Information-driven Science”, 2002, (http://www.umanitoba.ca/faculties/afs/plant_science/courses/39_769/lec01/lec01.1.html).
9. “Microarrays: Chipping away at the mysteries of science and medicine”, NCBI, 2003, (http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html).
10. DUGGAN D. J., BITTNER M., CHEN Y., MELTZER P., TRENT J. M., “Expression profiling using cDNA microarrays”, Nature genetics supplement, 1999, 21:10-14.
11. CHEN Y., DOUGHERTY E. R., BITTNER M., “Ratio-based decisions and the quantitative analisis of cDNA microarray images”, Journal of Biomedical Optics, 1997, 2(4):364-374.

12. BRAZMA A., VILO J., “Gene expression data analysis”, FEBS Letters, 2000, 480:17-24.
13. QUACKENBUSH J., “Computational analysis of microarray data”, Nature reviews genetics, 2001, 2:418-427.
14. EISEN M.B., SPELLMAN P.T., BROWN P.O., BOTSTEIN D., “Cluster analysis and display of genome-wide expression patterns”, Proc. Natl. Acad. Sci. USA, 1998, 95:14863-14868.
15. STURN A., ”Cluster analysis for large scale gene expression studies”, 2000, (http://genome.tugraz.at/Software/Genesis/Documentation.html).
16. DUTILH B. E., HOGEWEG P., “Gene networks from microarray data”, report Binf.1999.11.01. Bioinformatics, Utrecht University, 1999, (http://www-binf.bio.uu.nl/~dutilh/research/gene-networks).
17. BUHLER J., IDEKER T., HAYNOR D., “Dapple: Improved Techniques for Finding Spots on DNA Microarrays”, UW CSE Technical Report UWTR, 2000.
18. TANNER C., TEPESCH P., “GridGrinder Users Manual“, Corning inc., 2002.
19. EISEN M., “ScanAlyze User Manual”, Stanford University, 1998-9.
20. SHAROV V., “TIGR_Spotfinder 2.0 Overview”, (http://www.tigr.org/software/tm4/documentation.html).
21. JAIN A. N., TOKUYASU T., “Spot 2.0: User Manual”, UCSF Cancer Center, 2002.
22. EISEN M., “Cluster and TreeView Manual”, Stanford University 1998-9.
23. “Engene: User manual “ , (http://www.engene.cnb.uam.es/downloads/engeneUM.pdf).
24. STURN A., ”Genesis 1.0, Operation Manual”, IBMT-TUG, 2000.
25. “J-Express Help Files”, (http://www.ii.uib.no/~bjarted/jexpress/main.html).
26. LEMKIN P.F., ”MAExplorer reference manual”, National Cancer Institute, 2001.

27. SAEED A. I., BHAGABATI N., “TIGR MeV Multiexperiment Viewer”, TIGR, 2003.
28. IDEKER T., SCHWIKOWSKI B., “Cytoscape 1.1 Manual“, 2003.
29. BREITKREUTZ B., STARK C., TYERS M., “Osprey: a network visualization system”, Genome Biology, 2003, 4:R22.