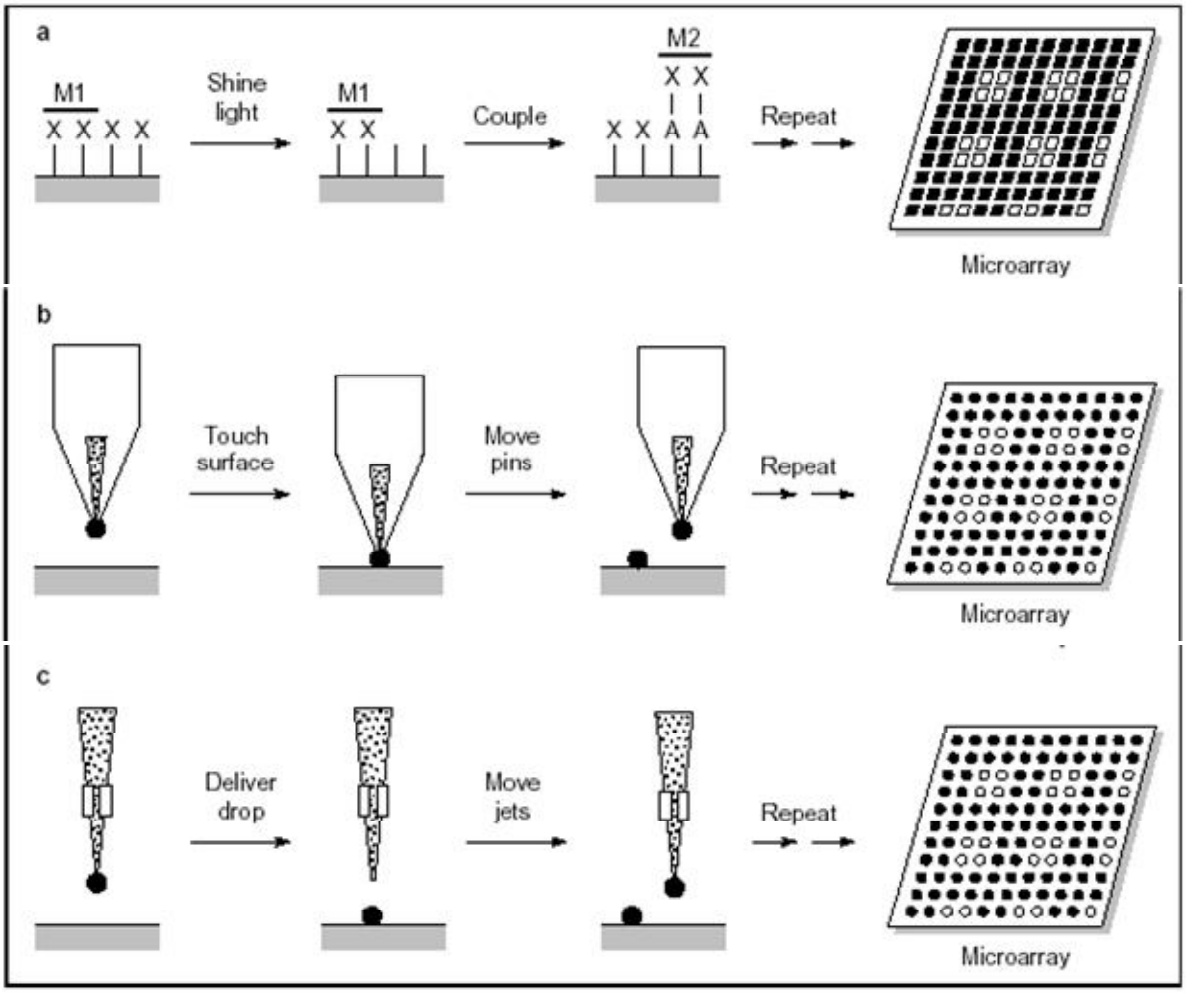


Dov Stekel: Bioinformatica dei Microarray - Capitoli 4-10 – copyright Stekel 2003
Tradotto e stampato su licenza della Cambridge University Press
Traduzione: Enrico D'Ascenzo – Revisione: Rodolfo Negri
Editing: Carlo Guerra

Capitolo 4

Image Processing

4.1 Introduzione

L'immagine del microarray generata dallo scanner (sezione 1.3) é costituita dai dati grezzi del tuo esperimento. Algoritmi del computer, conosciuti come software di estrazione degli spot, convertono l'immagine in informazione numerica che quantifica l'espressione del gene; questo è il primo passo dell'analisi dei dati. L'*image processing* che riguarda l'estrazione degli spot ha il maggiore impatto sulla qualità dei tuoi dati e sull'interpretazione che da essi puoi dedurre. Nel capitolo 1 abbiamo discusso tre tecnologie in accordo alle quali i microarrays sono costruiti: sintesi in *situ* con la piattaforma Affimetrix, arrays sintetizzati in *situ* mediante tecnologia a getto d'inchiostro, (Rosetta, Agilent e Oxford Gene Technology) e microarrays depositati. Affimetrix ha al suo interno algoritmi di *image processing* integrati nel processo sperimentale Gene Chip, in modo che l'utilizzatore finale non debba preoccuparsi di fare alcuna scelta al riguardo. Gli arrays a getto d'inchiostro sono di qualità molto piú elevata rispetto a quelli depositati e non soffrono della eventuale impraticità degli strumenti di *image processing* che gli spotted array richiedono; in aggiunta, Agilent fornisce il software per l'*image processing* adatto alla sua piattaforma, cosí che l'utilizzatore finale non debba preoccuparsi al riguardo.

Gli arrays depositati, di converso, forniscono all'utilizzatore finale un ampio range di strumenti con cui effettuare *l'image processing*. Queste scelte hanno un impatto sui dati, e perciò discuteremo in questo paragrafo i fondamenti di questi metodi computazionali, allo scopo di acquisire una migliore comprensione di come questi impattino sui dati.

4.2 Estrazione degli spots

Il primo passo nell'analisi computazionale dei dati del microarray consiste nel convertire le immagini digitali TIFF relative all'intensità di ibridizzazione, generate dallo scanner, in misure numeriche della intensità di ibridizzazione degli spot in ciascun canale. Questo processo è conosciuto come estrazione degli spot. Vi sono quattro passi:

1. Identificare gli spot sul microarray
2. Per ciascun spot, identificare i pixel che ne fanno parte.
3. Per ciascun spot, identificare quei pixel -posti nel contorno- che sono usati per il calcolo del background.
4. Calcolare numericamente informazioni sulla intensità degli spot, intensità del background e controllo di qualità del'informazione.

Discuteremo uno alla volta ciascuno di questi passi.

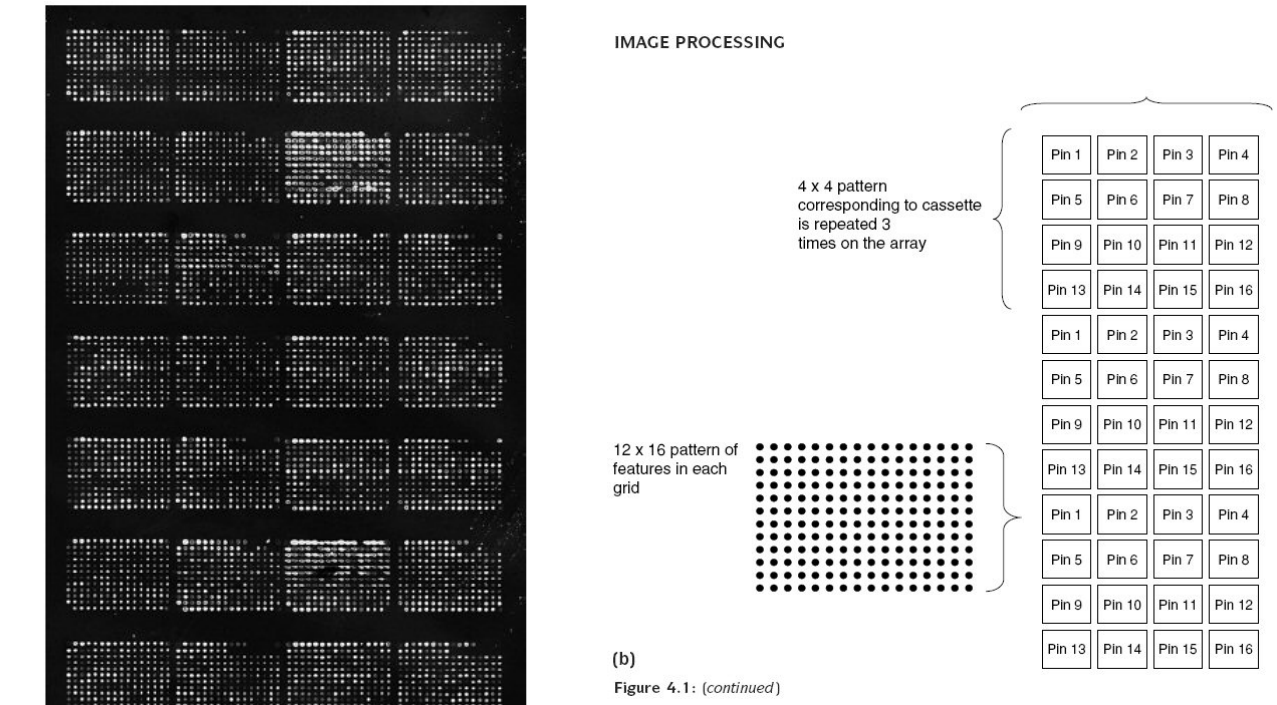


Figura 4.1: **(a)** Esempio di immagine di un microarray completo. In questo caso, vi sono 48 griglie organizzate in uno schema 12x4 e ciascuna griglia ha 12x16 spot. Pertanto, vi sono in totale 2.916 spot sull'array. **(b)** La distribuzione della griglia e degli spot in accordo con l'organizzazione dei puntali nel modulo. In questo esempio, lo spotting-robot ha 16 puntali distribuiti in uno schema 4x4. Ciascun puntale imprime tre griglie, in modo tale che lo schema di spotting dei 16 puntali è ripetuto tre volte sull'array. Ciascuna griglia è costituita da 192 spot distribuiti in uno schema 12x12. Il numero totale degli spot sull'array è 9216. Questi potrebbero essere tutti geni differenti (provenienti da 24 differenti piastre ciascuna delle quali ha 384 pozzetti), geni replicati in duplicato (12 differenti piastre a 384 pozzetti), oppure geni replicati in triplicato (8 differenti piastre a 384 pozzetti). *(prego vedere anche la sezione a colori alla metá del libro)*

Identificazione delle Posizioni degli Spots

Gli spots di molti microarrays sono organizzati secondo uno schema rettangolare. In generale, comunque, il pattern non è completamente regolare: gli spots sull'array sono distribuiti in griglie, con spazi maggiori tra le griglie che tra gli spots al'interno della griglia (Figura 4.1). Le griglie sono disposte in questo modo poiché vi sono parecchi puntali sul cassetto dello spotting robot; tutti gli spots di ciascuna griglia vengono impresse dallo stesso puntale.

Perché il software di estrazione degli spots funzioni, è necessario informare il sistema di quante griglie costituiscano l'array, insieme ai parametri associati con la griglia:

- Quante griglie ci sono in ciascuna direzione (x e y)
- Quanti spots per griglia ci sono in ciascuna direzione (x e y)
- Lo spazio tra griglia e griglia

Tutti i software di estrazione includono la possibilità di specificare questa informazione.

Esempio 4.1: Griglie dal “biorobotics microgrid spotting robot”

Uno Spotting Robot Microgrid Biorobotics è perfettamente allineato con una testina contenente 16 puntali secondo uno schema di 16x16. Esso è usato per fare un array con 9.216 spots (Figura 4.1) distribuiti in 48 griglie, ciascuna delle quali contiene 192 spots.

Ciascun puntale imprime (deposita) tre di queste griglie; queste potrebbero essere repliche dello stesso campione oppure potrebbero essere geni differenti.

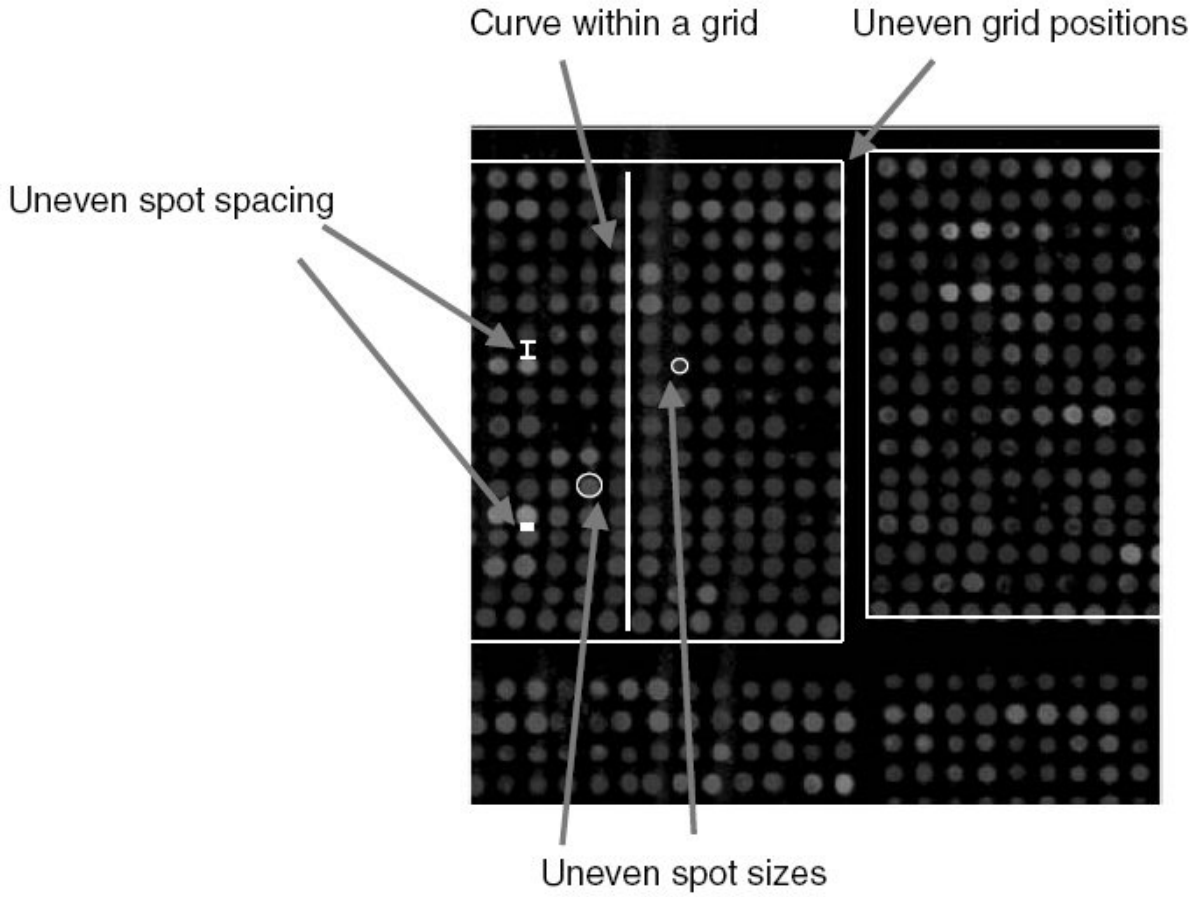


Figura 4.2: Problemi con le immagini di microarray con gli arrays depositati: **(1)** posizioni delle griglie distribuite non in modo regolare. Le due griglie non sono allineate. Questo succede poiché i puntali non sono perfettamente allineati nel modulo. **(2)** Curva dentro la griglia. Notare che i centri degli spots alla sommitá della linea verticale giacciono sulla linea, ma i centri degli spots sulla linea di base sono a sinistra della linea. Questo puó accadere se l’array non è orizzontale durante la costruzione, oppure a causa del movimento dei puntali durante la costruzione. **(3)** Spazio tra le griglie non regolare. Questo avviene a causa dei movimenti dei puntali durante la costruzione. Questo stesso fenomeno può accadere a causa del fatto che il vetrino non è perfettamente piatto. **(4)** Dimensioni degli spots non regolari. Diversi spots possono avere differenti dimensioni, e ciò può derivare da differenti quantità di liquido depositato sull’array. Questo può succedere anche se gli spots non si sono asciugati in modo regolare, perciò è importante mantenere costante la temperatura e l’umidità del’array durante il processo di costruzione *(vedere per favore anche il paragrafo a colori nella metá del libro)*

Un problema con l’identificazione delle posizioni degli spots sull’array sorge poiché le posizioni e le dimensioni degli spots dentro ciascuna griglia possono non essere uniformi (Figura 4.2). Vi sono almeno quattro difficoltà che possono insorgere:

- Posizione della griglia non regolare. Le griglie non sono allineate l’una con l’altra. Questo può succedere se i puntali non sono perfettamente allineati nel modulo.

- Curva dentro la griglia. Il vetrino non è perfettamente orizzontale, oppure i puntali si sono mossi, scivolando, nel modulo, e quindi gli spots sono impressi in un pattern curvo sulla superficie dell'array.
- Spazi fra gli spots non regolari. I puntali si sono mossi, scivolando, sul modulo, oppure la superficie del vetro non è perfettamente piatta.
- Dimensioni degli spots non regolari. Liquido depositato sul vetrino in modo irregolare durante la costruzione dell'array.

Tutti gli algoritmi software di estrazione degli spots trovano automaticamente la posizione di questi. Bisogna osservare, tuttavia, che nessuno di questi algoritmi è infallibile. La pratica corrente richiede una supervisione manuale del processo di estrazione degli spots per essere sicuri che tutti gli spots siano identificati dal software; di solito qualche intervento manuale è richiesto per allineare il software cosí che gli spots siano identificati. La maggior parte e dei pacchetti di software di estrazione degli spots hanno queste proprietà.

Identificazione dei Pixel che comprendono gli spots

Il prossimo passo nella procedura di estrazione degli spots è chiamato segmentazione; questo è il processo con cui il software determina quali pixel - fra quelli che si trovano nel'area di uno spot - appartengano allo spot stesso, così che la loro intensità abbia rilevanza ai fini di una misura quantitativa della intensità di quello spot. Vi sono quattro metodi di segmentazione comunemente usati:

- Cerchio fisso
- Cerchio variabile
- Istogramma
- Forma adattiva

Diversi pacchetti di software implementano differenti algoritmi di segmentazione (Tabella 4.1) e qualche pacchetto implementa più di un algoritmo sulla stessa immagine.

TABLE 4.1: Segmentation Algorithms of Common Image-Processing Software Packages

Segmentation Method	Software Implementing Method
Fixed Circle	ScanAnalyze GenePix QuantArray
Variable Circle	QuantArray GenePix Dapple Agilent Feature Extraction
Histogram	ImaGene QuantArray
Adaptive Shape	Spot

Segmentazione con Cerchio fisso

La segmentazione con cerchio fisso pone un cerchio di dimensione determinata sopra lo spot e considera tutti i pixel all'interno del cerchio come facenti parte dello spot. (Figura 4.3a). Il problema con la segmentazione a cerchio fisso è che essa fornisce risultati poco accurati se gli spots sono di differenti dimensioni - che è di solito il caso quando si lavora con molti microarrays. Pertanto, la segmentazione a cerchio fisso dovrebbe essere evitata il più possibile.

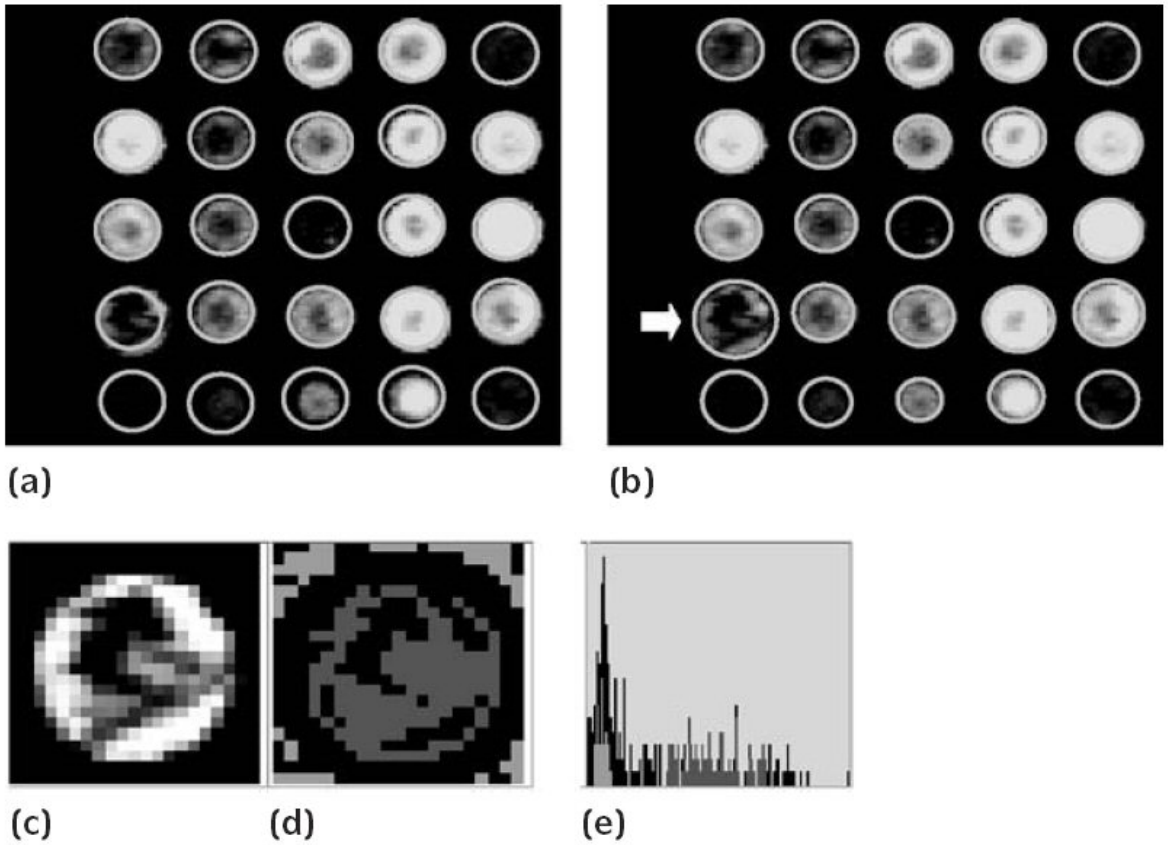


Figura 4.3: **(a)** Segmentazione con cerchio fisso. Un cerchio della stessa dimensione è posto su ogni spot dell'array ed i pixel dentro il cerchio sono utilizzati per determinare l'intensità dello spot, Questo non è un buon metodo poiché il cerchio potrebbe essere troppo grande per alcuni spots, mentre per altre potrebbe essere troppo piccolo. **(b)** Segmentazione con cerchio variabile. Un cerchio di differenti dimensioni è applicato su ciascun spot ed i pixel dentro il cerchio sono usati per determinare l'intensità dello spot. Questo metodo funziona meglio quando si abbia a che fare con spots di differenti dimensioni, ma non funziona così bene quando gli spots abbiano differente forma, per esempio lo spot irregolare, di colore rosso, contrassegnato con una freccia. **(c)** Zoom sul canale rosso dello spot dalla forma irregolare, marcato con una freccia in (b). Notare che nella regione nera non vi è alcuna ibridazione, probabilmente perché non vi è nessuna sonda legata al vetrino in quella area. **(d)** metodo dell'istogramma applicato a quello spot. I pixel rossi sono quelli che sono stati usati per calcolare il segnale dello spot; i pixel verdi sono stati usati per calcolare il background dello spot. I pixel neri non sono stati usati affatto. L'area corrispondente alla regione nera (c) non è usata per calcolare l'intensità dello spot. Anche i pixel più luminosi sono stati scartati. Il rapporto rosso/verde di questo spot, calcolato con il metodo di segmentazione a cerchio fisso è uguale ad 1.8, mentre con il metodo a cerchio variabile è 1.9, ed infine con il metodo dell'istogramma è 2.6; la differenza di espressione del gene che misuriamo nei vari campioni è diversa a seconda dell'algoritmo che si usa. A causa della forma irregolare dello spot, il metodo ad istogrammi è probabilmente quello che fornisce i migliori risultati. **(e)** Istogramma delle intensità dei pixel in uno spot di forma irregolare. Le barre rosse rappresentano i pixel usati per l'intensità del segnale; le barre verdi rappresentano i pixel usati per l'intensità del background, le barre nere sono pixel non usati per la quantizzazione. I pixel più luminosi ed i pixel più scuri non vengono usati; in questo modo si perviene ad una migliore misura dell'intensità di ibridazione.

Segmentazione con Cerchio Variabile

La segmentazione con cerchio variabile adatta un cerchio di dimensione variabile sulla regione contenente lo spot (Figura 4.3b). Questo metodo si presta bene a risolvere spot di dimensioni variabili, ma lavora meno bene con spot di forma irregolare.

Segmentazione con Istogramma

La segmentazione con istogramma adatta un cerchio sia al di sopra la regione dello spot che sul background e quindi valuta un istogramma dell'intensità dei pixel nello spot (Figura 4.3e). I pixel più luminosi e quelli più deboli non vengono utilizzati nella quantificazione dell'intensità dello spot. La segmentazione ad istogrammi fornisce risultati affidabili per spot di forma irregolare. I metodi ad istogrammi possono essere instabili per spots piccoli se la maschera circolare è molto grande.

Segmentazione con Forma Adattiva

Questa consiste in un algoritmo molto sofisticato che può lavorare bene anche con spots di forma irregolare. L'algoritmo richiede un piccolo numero di *pixel seme* nel centro di ciascuno spot per iniziare l'elaborazione. Quindi, esso estende la regione di ciascuna feature aggiungendo pixel che sono simili in intensità ai pixel delle zone di contorno.

Esempio 4.2: Estrazione dello spot quando questo sia di forma irregolare

Il programma *QuantArray* puó utilizzare i metodi di segmentazione a cerchio fisso, a cerchio variabile e ad istogramma per determinare l'intensità degli spots. Esso è stato utilizzato per uno spot di forma irregolare indicato con una freccia in Figura 4.3b. Il rapporto di intensità della luce rossa rispetto alla luce verde per ciascuno dei metodi su elencati è il seguente:

- Cerchio Fisso: 1.8
- Cerchio Variabile: 1.9
- Istogramma: 2.6

In questo caso, la segmentazione ad istogramma sembra fornire i risultati piú affidabili poiché l'area dello spot non ibridizzato non è inclusa. In questa regione, è probabile che non vi sia nessuna molecola della sonda legata all'array.

Identificazione dei Pixel di Background

L'intensità del segnale degli spots include i contributi provenienti dalle ibridizzazioni non specifiche e di altre fluorescenze provenienti dal vetrino. Di solito si determina questa fluorescenza calcolando il segnale di background dai pixel che sono vicini a ciascun spot ma che non ne fanno parte. Differenti pacchetti di software usano differenti regioni vicino a ciascun spot come pixel di background (Figura 4.4). L'intensità di background é sottratta dall'intensità dello spot per fornire una stima più affidabile dell'intensità di ibridazione di ciascun spot. La sottrazione del background è discussa con maggior dettaglio nel paragrafo 5.2.

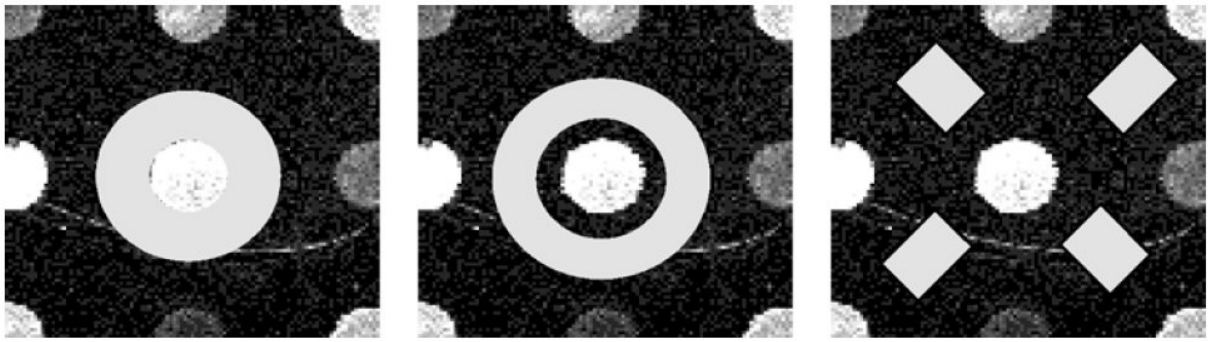


Figura 4.4: Regioni di background usati da differenti software. Pacchetti software diversi usano differenti regioni di pixel intorno allo spot per determinare l'intensità di background. **(a)** ScanAlyze: la regione è adiacente allo spot. Questo sarebbe non accurato se lo spot fosse piú grande della dimensione fissata del circolo usato per la segmentazione. **(b)** ImaGene: c'è uno spazio tra lo spot ed il background: questo è un metodo migliore rispetto ad (a). **(c)** Spot e GenePix: la regione di background è nel'intermezzo tra gli spots: anche questo è un buon metodo.

Calcolo dell'Informazione Numerica

Avendo determinato i pixel rappresentanti ciascun spot, il software di *image-processing* deve calcolare l'intensità di ciascuno di essi. Il software di *Image-processing* fornisce tipicamente un certo numero di misure:

- Media del Segnale: la media dei pixel che appartengono allo spot
- Media del Background: la media dei pixel che appartengono al background intorno allo spot
- Mediana del Segnale: la mediana dei pixel che appartengono allo spot
- Mediana del Background: la mediana dei pixel appartenenti al background
- Deviazione Standard del Segnale: la deviazione standard dei pixel appartenenti allo spot
- Deviazione standard del Background: deviazione standard dei pixel appartenenti al background
- Diametro: il numero di pixel compresi nella larghezza dello spot
- Numero di Pixel: il numero di pixel appartenenti allo spot
- Flag: una variabile che assume il valore 0 se lo spot è buono, e qualunque altro valore se lo spot non è buono.

La tabella 4.2 mostra qualche esempio dei dati che escono da *ImaGene*, che usa alcuni di questi campi. Vi sono un certo numero di modi in cui questa informazione può essere usata. Il dato più importante è la misura dell'intensità di ibridazione di ciascun spot.

Qui l'utilizzatore deve effettuare una scelta tra media e mediana dell'intensità dei pixel. In generale è preferibile utilizzare la mediana anziché la media. La ragione di ciò è che la mediana è più robusta nel delineare i pixel rispetto alla media: un piccolo numero di pixel molto luminosi (emergenti dal rumore) hanno la capacità di far degenerare la media, ma lasciano abbastanza immutata la mediana.

TABLE 4.2: Example Output from ImaGene

Meta Row	Meta Column	Row	Column	Gene ID	Flag	Signal Mean	Background Mean	Signal Median	Background Median	Signal Stdev	Background Stdev	Diameter
1	1	1	1	H3126A06-3	0	29453.02	99.27228	31434	94.5	5165.786	31.88765	14
1	1	1	2	H3126A06-3	0	35591.85	134.6042	35718.5	124	3781.532	54.10425	14
1	1	1	3	H3126C06-3	0	455.3077	109.5556	436	109	112.4157	20.80846	14
1	1	1	4	H3126C06-3	0	780.3725	95.34177	786	95	136.1061	14.2317	14

Nota: I primi quattro spots da un array in cui gli spots sono stati estratti con ImageGene. Le prime cinque colonne forniscono informazione circa gli spots stessi: la griglia, la posizione della griglia, e l'ID del gene. La successiva colonna è il flag, che è zero se lo spot è buono, e diverso da zero se vi è qualche problema con lo spot. Le due colonne successive sono la media dei pixel comprendente, rispettivamente, lo spot ed il background, seguite da due colonne per la media di questi pixel e due colonne per la deviazione standard degli stessi. La colonna finale rappresenta il numero di pixel inter-spots come misura della dimensione dello spot. Dal diametro, si può determinare in modo approssimato il numero totale dei pixel dello spot, in questo caso circa 150.

Esempio 4.3: Spot con polvere luminosa

Lo spot mostrato in Figura 4.5 ha un zona di polvere luminosa all'interno. Il rapporto rosso su verde, usando la media dell'intensità dei pixel, è uguale ad 1.9 lo stesso rapporto usando la mediana dell'intensità dei pixel è 1.3. La media è variata a causa dei pixel luminosi della polvere, e quindi il risultato è molto differente che nel caso della mediana.

La mediana è una misura più robusta. Questo fatto può essere verificato rimuovendo la polvere sull'immagine e ricalcolando le intensità: la media è 1.4 e la mediana è 1.3.

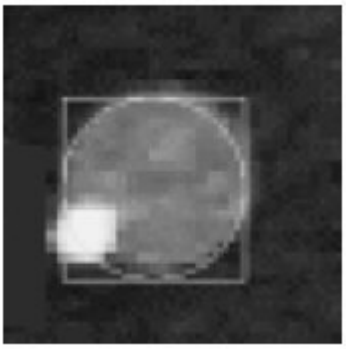


Figura 4.5: Spot con polvere luminosa. Uno spot altrimenti accettabile ha una zona luminosa dovuta alla polvere depositata su di esso. In questo caso, la media delle intensità dei pixel costituirà una misura inaffidabile dell'intensità dello spot, poiché essa sarà "spostata" dalla luminosità dei pixel, mentre la mediana sarà una misura robusta rispetto al rumore.

La seconda ed importante informazione numerica è la deviazione standard del segnale. Questa è usata come controllo di qualità dell'array in due modi differenti (Figura 4.6):

- Come misura del controllo di qualità dello spot. Se la deviazione standard dello spot è maggiore, diciamo, del 50% dell'intensità della mediana, lo spot potrebbe essere scartato.
- Per determinare se l'array è in condizioni di saturazione. Il problema di saturazione degli spot consiste nel fatto che non conosciamo la vera intensità dello spot saturo, e così non è possibile usare tale spot come parte di un'analisi quantitativa dell'espressione del gene.

La terza informazione importante proveniente dal software di estrazione degli spots è *il flag*. Questo è uguale a 0 se lo spot è buono, ma sarà diverso da 0 se lo spot ha qualche problema. Differenti software di image-processing utilizzano differenti valori del flag per differenti problemi, ma i problemi tipici sono:

- Spot pessimo: La deviazione standard dei pixel è molto piú alta in relazione alla media dei pixel.
- Spot negativo: Il segnale dello spot è inferiore al segnale di background.
- Spot scuri: Il segnale di questi spots è molto debole.
- Spots il cui flag è stato impostato manualmente. Il ricercatore ha posto il flag sullo spot utilizzando il software *image-processing*.

È anche usuale la pratica di rimuovere il flag dagli spot per una successiva analisi.

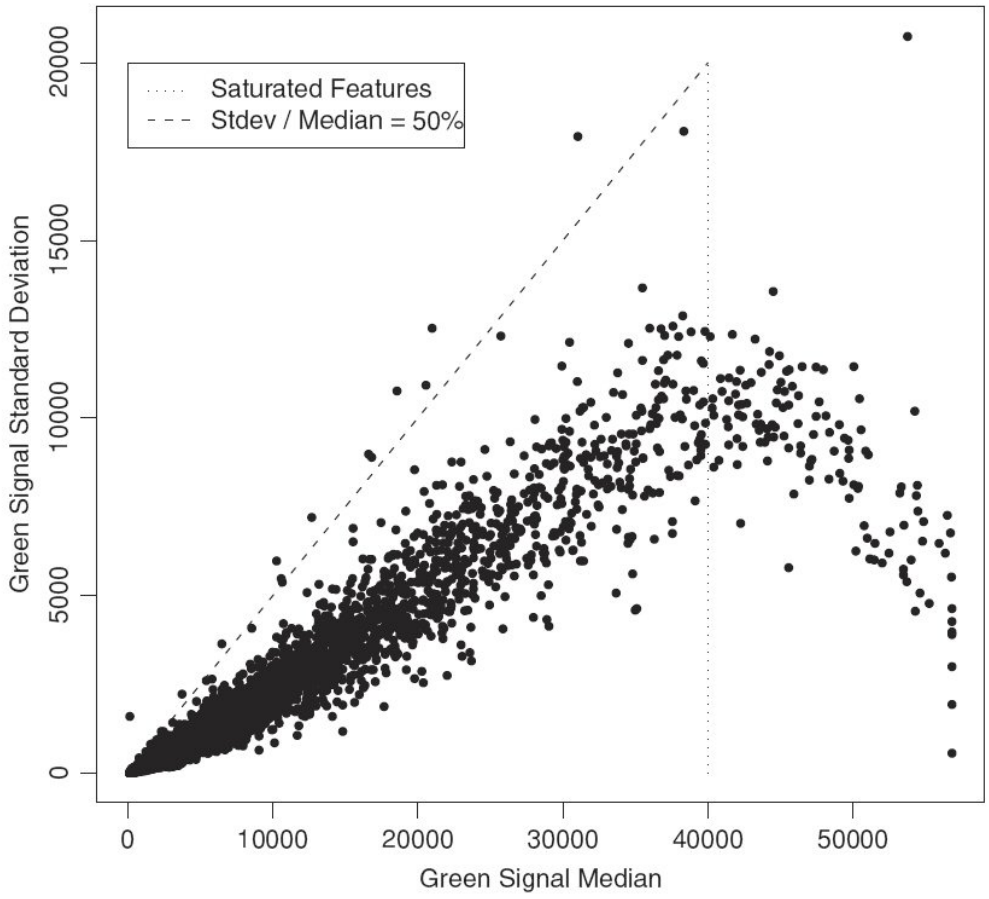


Figura 4.6: Diagramma dell'uso della mediana versus la deviazione standard. La deviazione standard dell'intensità dei pixel per il segnale Cy3 (verde) degli spots di un array è rappresentata graficamente in funzione della mediana delle intensità dei pixel come metodo per il controllo di qualità. Per spots che hanno un'intensità del segnale inferiore a 40.000, la deviazione standard è approssimativamente proporzionale alla media, con un coefficiente di variabilità (Capitolo 6) di circa il 23%. Tuttavia, per spots che presentano un'alta intensità del segnale, la deviazione standard diminuisce. Questo accade poiché questi spots hanno pixel saturati che presentano la stessa intensità, (con questo particolare scanner, la massima intensità dei pixel è 56.818). Gli spots molto luminosi hanno tutti i pixel saturati, così che la loro deviazione standard è uguale a 0. Gli spots saturati non possono essere usati per l'analisi quantitativa dell'espressione differenziale del gene. Il diagramma mostra anche gli spots - da scartare - con deviazione standard molto alta. Questi spots potrebbero essere inaffidabili e dovrebbero essere escluse da analisi successive. La regione triangolare indicata rappresenta gli spots affidabili che non sono saturati e che hanno un coefficiente di variabilità inferiore al 50%. Alcuni spots situati nella parte alta del triangolo potrebbero essere ritenuti non buoni (e quindi da scartare), e dunque esclusi dall'analisi. (Questi dati sono stati ottenuti in modo privato da Ed Southern.)

Riassunto dei punti chiave

- L’immagine nativa del tuo array è costituita dai dati grezzi.
- Il software di estrazione degli spots calcola le misure numeriche dell’espressione del gene dall’immagine.
- La scelta degli algoritmi di estrazione degli spots avrà un impatto consistente sui dati che hai generato.

Capitolo 5

Normalizzazione

5.1 Introduzione

La normalizzazione è il nome generale con cui vengono denominati un insieme di metodi che hanno lo scopo di trovare la soluzione agli errori sistematici ed agli errori di polarizzazione introdotti dalla piattaforma sperimentale di microarray. I metodi di normalizzazione contrastano con i metodi di analisi dei dati descritti nei Capitoli 7, 8 e 9 che sono utilizzati, quest'ultimi, per fornire una risposta scientifica alle questioni per cui l'esperimento a microarray è stato sviluppato. Lo spirito di questo capitolo è di fornire una comprensione del perché necessitiamo di normalizzare i dati del microarray, e di illustrare quei metodi di normalizzazione che sono più comunemente usati. Questo capitolo si divide nei successivi tre paragrafi:

- Paragrafo 5.2: Pulizia e Trasformazione dei dati: vengono esaminate le prime fasi della pulizia e della trasformazione dei dati generati dal software di estrazione degli spots, prima che ogni ulteriore analisi venga sviluppata.
- Paragrafo 5.3: Normalizzazione all'interno dell'array, vengono trattati i metodi che permettono il confronto dei canali dei fluorocromi Cy3 e Cy5 di un microarray a due colori.
- Paragrafo 5.4: Normalizzazione tra gli arrays, vengono descritti i metodi che permettono il confronto di misure su differenti arrays. Questo paragrafo è applicabile sia agli arrays a due colori, sia a quelli a singolo canale, inclusi gli arrays Affimetrix.

5.2 Pulizia dei dati e trasformazione

I dati di un microarray generati dal software di estrazione degli spots sono tipicamente nella forma di files di testo (Tabella 4.2). Prima di utilizzare i dati per rispondere a questioni scientifiche, vi sono un certo numero di passi che vengono comunemente intrapresi per assicurarsi che i dati stessi siano di alta qualità e siano adatti all'analisi successiva.

Questo paragrafo descrive le tre fasi relative alla pulizia dei dati e alla trasformazione di questi.

- Rimozione degli spots marcati con il flag
- Sottrazione del background
- Utilizzo dei logaritmi

Rimozione degli spots marcati con il flag

Nel Capitolo 4 abbiamo descritto quattro tipi di spots marcati con il flag: spots non buoni, spots negativi, spots scuri e spots il cui flag è stato impostato manualmente. Questi sono spots per i quali il software di image-processing ha rivelato un qualche tipo di

problema. Il primo – che é di gran lunga l'approccio piú comune- consiste nel rimuovere dall'insieme dei dati quegli spots marcati con il flag. Questo processo è immediato, ma presenta lo svantaggio che alcune volte potrebbero essere rimossi dati potenzialmente validi. Il secondo metodo è quello di procedere all'indietro e risalire all'immagine originale di ogni spot marcato con il flag, ed identificare il problema che ha fatto sí che quello spot venisse marcato. Se risulta conveniente, lo sperimentatore può sviluppare l'operazione di estrazione di un nuovo spot ed ottenere una misura più affidabile dell'intensità del segnale.

Questo procedimento ha lo svantaggio di richiedere tempo e risorse, e può non sempre essere pratico.

Sottrazione del background

Il secondo passo nella pulizia dei dati del microarray consiste nel togliere il segnale di background dalla intensità dello spot. Si pensa che Il segnale di background sia dovuto al contributo di ibridizzazioni non specifiche dei target marcati sul vetrino, oltre che alla naturale fluorescenza del vetrino stesso. Questo procedimento funziona bene quando l'intensità dello spot è più alta dell'intensità del background. Di converso, quando l'intensità del background è più elevata rispetto all'intensità dello spot, il risultato dovrebbe dare un numero negativo, che non avrebbe alcun senso. Vi sono tre approcci per gestire questa situazione:

- Rimuovere questi spots dall'analisi. Poiché l'intensità dello spot dovrebbe essere piú elevata rispetto all'intensità del background, un valore insolitamente elevato del background è indice di un problema locale con l'array e quindi l'intensità dello spot è considerata inaffidabile. Questo è un approccio molto comune.
- Sostituire l'intensità –alla quale è stato sottratto il background- con il valore minimo. Questo, tipicamente, assumerà il valore 1[1]. L'idea sulla quale poggia questa congettura é che se l'intensità del background é più alta della intensità dello spot, essa rappresenta un gene con espressione nulla -o con espressione molto bassa, e quindi viene usato il valore più basso disponibile.
- Uso di algoritmi più sofisticati (Bayesiani) per stimare la vera intensità dello spot, basato sull'assunzione che l'intensità vera dello spot è più alta della intensità del background, e quindi un elevato valore di background rappresenta un qualche tipo di errore sperimentale[2].

Dati Affimetrix

I dati provenienti da Affimetrix *GeneChip* possono essere affetti da un problema similare, e richiedono, quindi, un approccio adeguato perché il problema venga risolto.

L'espressione del gene viene determinata confrontando l'intensità del segnale dovuto alla ibridizzazione alle sonde complementari al gene che viene misurato, con l'intensità del segnale dovuto alla ibridizzazione alle sonde che contengono sequenze imperfette (contenenti variazioni n.d.t.); il segnale proveniente dalle sonde imperfette si ritiene dovuto a cross-ibridizzazione. Nelle prime versioni del software Affimetrix (versione 4 ed inferiori), l'espressione del gene veniva calcolata come una combinazione lineare delle differenze tra le sonde vere e le sonde disadattate. Quando le sonde imperfette hanno una intensità

[1] Le letture da uno scanner a 16 bit sono segnali digitali e quindi sono numeri interi; lo zero non può essere utilizzato poiché è pratica comune utilizzare il logaritmo del segnale nel prossimo passo di analisi.
[2] Un riferimento che entra nel dettaglio di tale metodo é dato alla fine del capitolo.

del segnale più alta rispetto alle sonde complementari, il software genera numeri negativi, che non sono particolarmente significativi. Vi sono quattro possibili approcci per gestire i geni con intensità negative:

- Scartare questi geni dall'analisi. Il ragionamento seguito è che se le sonde imperfette hanno un segnale più alto delle sonde complementari, il segnale è soprattutto cross-ibridizzazione, e quindi è inaffidabile.
- Sostituire i numeri negativi con numeri positivi di ampiezza minima, di solito con 1. Il ragionamento che sta alla base di questa congettura è che i geni per i quali il segnale delle sonde imperfette è inferiore a quello delle sonde perfette sono *non espressi,* oppure espressi ad un livello molto basso, e quindi sostituiamo il segnale con il più basso valore possibile.
- Uso di algoritmi più sofisticati per stimare il valore vero della intensità dello spot, basandoci sull'assunzione che il valore vero dell'intensità della sonda sia piú alto dell'intensità del segnale imperfetto; pertanto, l'effetto è un artefatto e rappresenta un qualche tipo di errore sperimentale.[3]
- Affimetrix ha cambiato il suo algoritmo nella versione 5 del software cosí che non è più possibile che vengano fuori risultati negativi. Pertanto, questo problema è presente soltanto nei dati storici di Affimetrix che non sono stati ri-analizzati con il loro ultimo software.

Introduzione dei Logaritmi

È pratica comune trasformare i dati di intensità del microarray a DNA dalla loro forma sorgente (n.d.t. forma lineare con cui escono dal Convertitore Analogico/Digitale dello scanner), alla forma logaritmica prima di procedere con l'analisi[4]. Questa trasformazione si pone parecchi obiettivi, e cioè:

- Dovrebbe esserci una ragionevole distribuzione dei valori su tutto l'intervallo di intensità.
- La variabilità dovrebbe essere costante per tutti i livelli di intensità.
- La distribuzione degli errori sperimentali dovrebbe essere approssimativamente normale.
- La distribuzione delle intensità dovrebbe essere approssimativamente ben conformata (n.d.t.: per curva "ben conformata" in matematica si intende una curva con variazioni molto lente della monotonicità, rispetto ad un modello).

È usuale usare il logaritmo in base 2 nell'analisi dei dati dei microarrays. La ragione è che il rapporto delle intensità sorgenti dei canali di fluorocromi Cy5 e Cy3 è trasformato in differenze tra i logaritmi delle intensità dei canali di Cy5 e Cy3. Pertanto, geni due volte sovraespressi rispetto al controllo corrisponderanno ad un logaritmo del rapporto = +1 e quelli due volte sottoespressi ad un logaritmo del rapporto = -1.

[3] Al momento in cui questo libro é stato scritto, io non ero a conoscenza di alcun metodo pubblicato per sviluppare tale analisi con i dati Affimetrix. Di converso, la metodologia Bayesiana applicata alla sottrazione del background potrebbe essere modificata ed applicata anche a questo problema.

[4] Vi é un certo numero di trasformazioni alternative che possono essere applicate al posto dei logaritmi, tipicamente motivati dallo spirito di ottenere che la variabilità sia costante per tutti i livelli di intensità. Un riferimento a questi metodi è riportato alla fine di questo capitolo.

TABLE 5.1: Conversion from Log (to Base 2) to Raw Intensity and from Raw Intensity to Log (to Base 2) Intensity

Log (to Base 2) Intensity	Raw Intensity	Raw Intensity	Log (to Base 2) Intensity
0	1	1	0
1	2	2	1
2	4	5	2.32
3	8	10	3.32
4	16	20	4.32
5	32	50	5.64
6	64	100	6.64
7	128	200	7.64
8	256	500	8.97
9	512	1,000	9.97
10	1,024	2,000	10.97
11	2,048	5,000	12.29
12	4,096	10,000	13.29
13	8,192	20,000	14.29
14	16,384	50,000	15.61
15	32,768		

Geni che non sono differenzialmente espressi avranno un logaritmo del rapporto = 0. Questi logaritmi del rapporto hanno una simmetria naturale che riflette la biologia e non è presente nei rapporti dei dati grezzi. Ad esempio i geni 2 volte sovraespressi, non differenzialmente espressi o due volte sottoespressi danno rapporti grezzi di 1, 0 e 0.5, rispettivamente.

I logaritmi in base 2 in un certo range delle intensità sono mostrati in tabella 5.2.

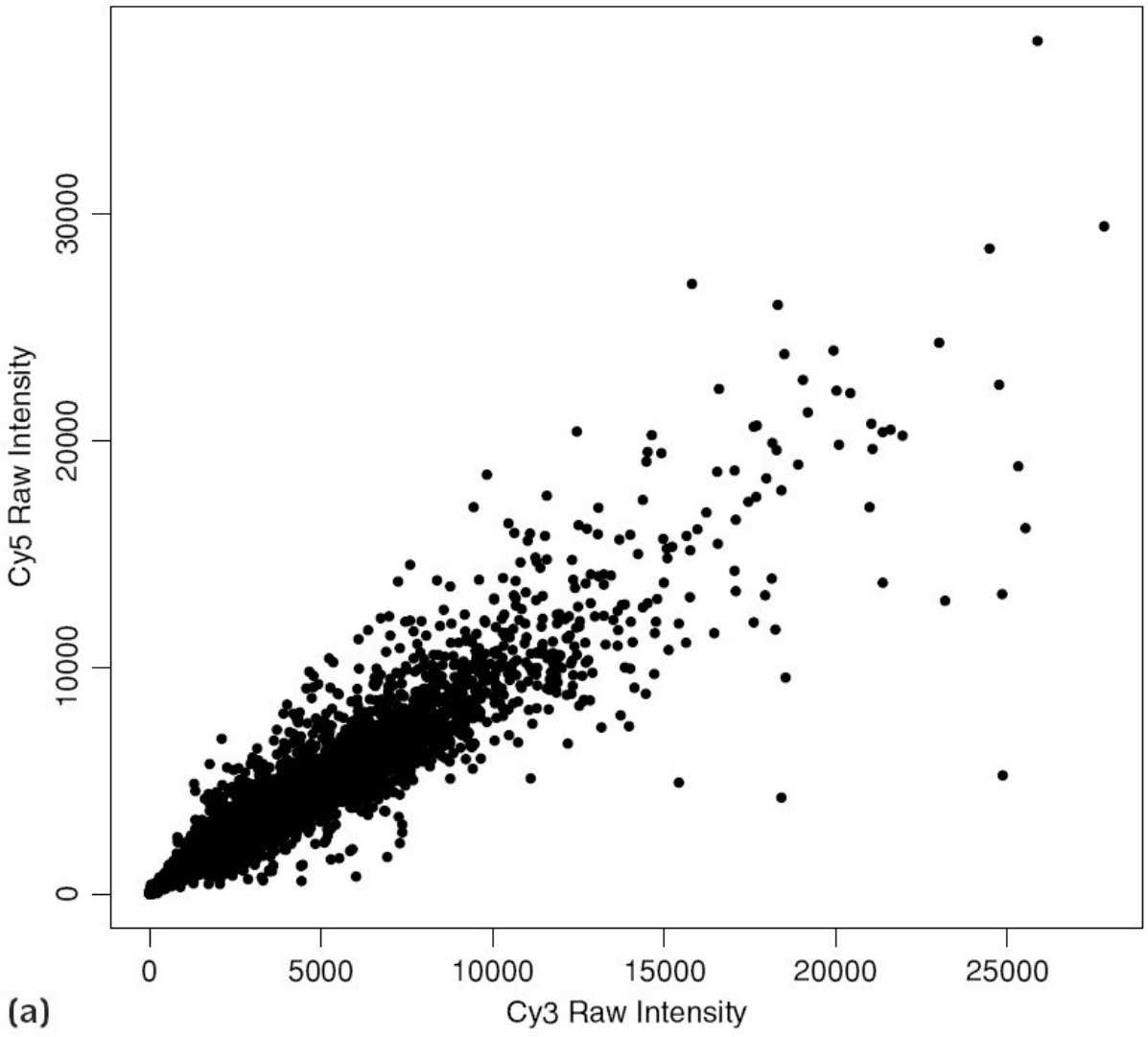
Il logaritmo in base 2 è strettamente correlato al logaritmo naturale preferito dai matematici. I logaritmi naturali sono implementati in Excel, R e nei calcolatori da taschino. Si può convertire un logaritmo naturale in un logaritmo in base 2 con la seguente equazione:

$$\log(\text{to base } 2)x = \frac{\log(\text{natural})x}{\log(\text{natural})\ 2}$$

Equazione 5.1

TABLE 5.2: Conversion from Fold Ratios to Log (to Base 2) Ratios

Fold Ratio	Log (to Base 2) Ratio Difference
4-fold down-regulated	−2
3-fold down-regulated	−1.58
2-fold down-regulated	−1
1.5-fold down-regulated	−0.58
No change	0
1.5-fold up-regulated	0.58
2-fold up-regulated	1
3-fold up-regulated	1.58
4-fold up-regulated	2



(a)

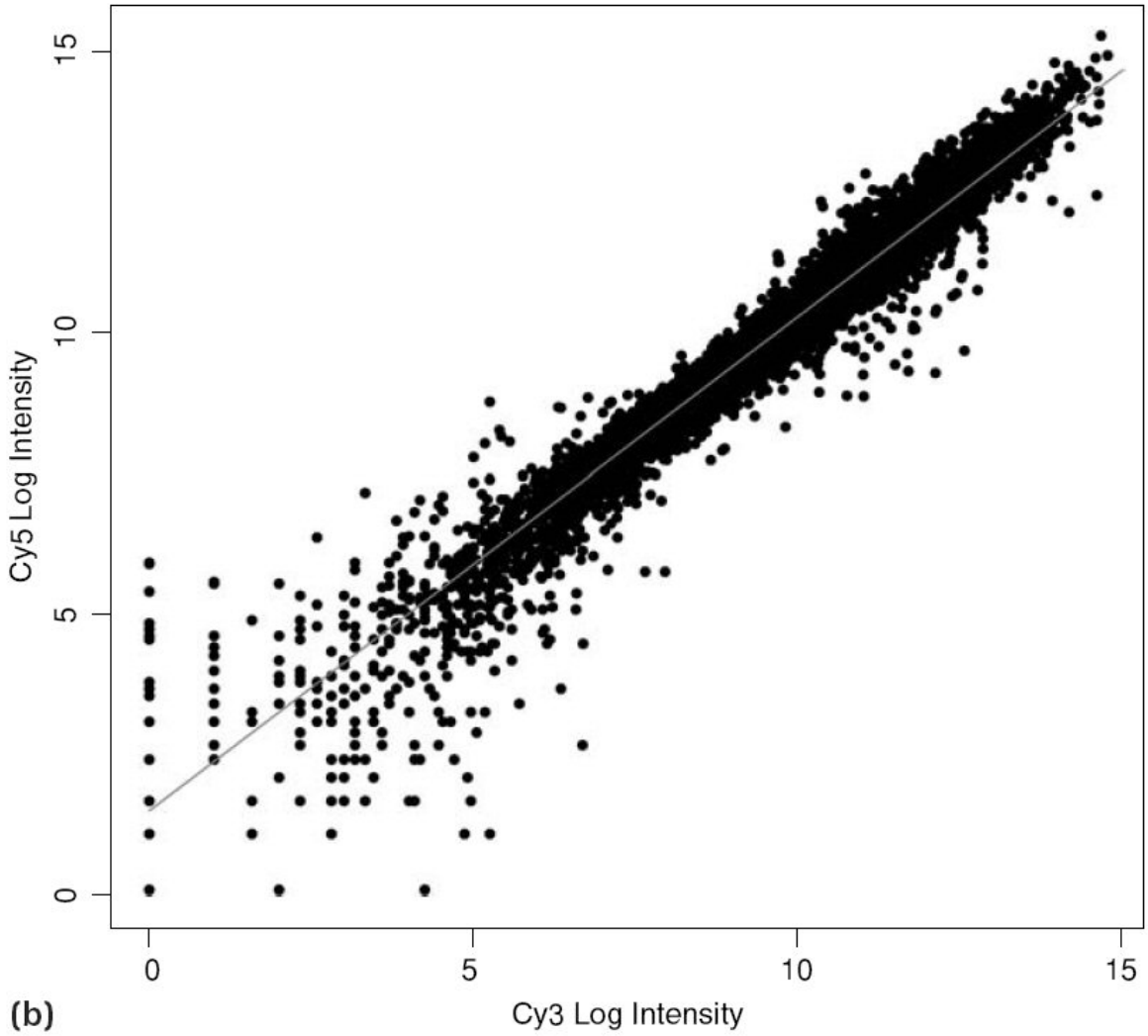


Figura 5.1: Grafici del fluorocromo Cy3 in funzione del fluorocromo Cy5 per l'insieme dei dati 5A. I fibroblasti umani del prepuzio sono stati infettati con *Toxoplasma gonadii* per un periodo di 1 ora. È stato preparato un campione, marcato con fluorocromo Cy5 (rosso) ed ibridizzato ad un microarray con, approssimativamente, 23.000 spots. Il canale del fluorocromo Cy3 (verde), è un campione preparato da fibroblasti non infettati. Siccome il periodo di infezione è breve, molti geni in questo esperimento non sono differenzialmente espressi.

(a) Diagramma di dispersione delle intensità sorgenti (background sottratto); ciascun punto nel grafico rappresenta uno spot sull'array, con la coordinata x rappresentante l'intensità Cy3, e la coordinata y rappresentante l'intensità Cy5. Il grafico dimostra due inadeguatezze dei dati sorgenti, che avrebbero -se usati *sic et simpliciter*- un impatto negativo sulla successiva analisi dei dati.

1. La maggior parte dei dati sono agglomerati sull'angolo a sinistra in basso, con relativamente pochi dati lungo il diagramma.
2. La variabilità dei dati aumenta con l'intensità, così che essa è molto bassa quando l'intensità è bassa, ed è molto elevata quando l'intensità é elevata.

(b) Grafico di dispersione del logaritmo delle intensità (in base 2). Questo diagramma è migliore di quello in (a). I dati sono distribuiti sull'intero intervallo delle intensità, e la variabilità dei dati è la stessa alla maggior parte delle intensità. I geni con una intensità logaritmica inferiore a 5 presentano una variabilità leggermente una più alta, ma questi geni sono molto poco espressi e sono al di sotto del livello di rivelazione della tecnologia del microarray. La linea retta è la regressione lineare calcolata attraverso i dati. La regressione lineare non è perfetta (i dati sembrano che curvino verso l'alto della linea alle intensità più alte), ma essa si puó considerare approssimativamente una retta. L'intercetta è 1.4, ed il gradiente è 0.88. Se i due canali si comportassero in modo identico, l'intercetta sarebbe 0 ed il gradiente sarebbe 1. Possiamo concludere che i due canali di fluorocromi Cy3 e Cy5 si comportano in modo differente alle differenti intensità; ciò potrebbe risultare da differente incorporazione del fluorocromo oppure da differenti risposte dei fluorocromi ai laser.

Esempio 5.1: Utilizzo del log del toxoplasma gonadii (dataset 5a)

I fibroblasti presi dal prepuzio umano sono stati infettati con *TOXOPLASMA GONADII*. I campioni provenienti dalle celle non infettate e quelli trattati con *T. gonadi* per 1 ora sono ibridizzati a due canali di microarray con, approssimativamente, 23000 spots[5]. I ricercatori vogliono identificare i geni che sono espressi in modo differenziale. I dati grezzi (Figura 5.1a) non soddisfano i requisiti per una analisi efficace. La maggior parte degli spots sono nella parte bassa a sinistra del grafico; la variabilità aumenta con l'intensità, e la distribuzione della intensità non è conformata come una campana, ma è fortemente spostata a destra (Figura 5.2a e 5.2b). I dati logaritmici (Figura 5.1b), tuttavia, soddisfano i requisiti. I dati sono ben distribuiti lungo l'intervallo dei valori del logaritmo dell'intensità, la variabilità è approssimativamente costante a tutte le intensità e sembrerebbe essere normalmente distribuita (ad eccezione di quei geni che sono espressi molto debolmente, le cui intensità sono probabilmente inaffidabili), e la distribuzione di intensità (Figure 5.2c e 5.2d) sono molto vicine alla forma a campana (benché queste distribuzioni siano leggermente spostate a destra.

[5] L'articolo dal quale questi dati sono stati mutuati é citato alla fine del capitolo. I dati sono disponibili presso lo Stanford Microarray Database

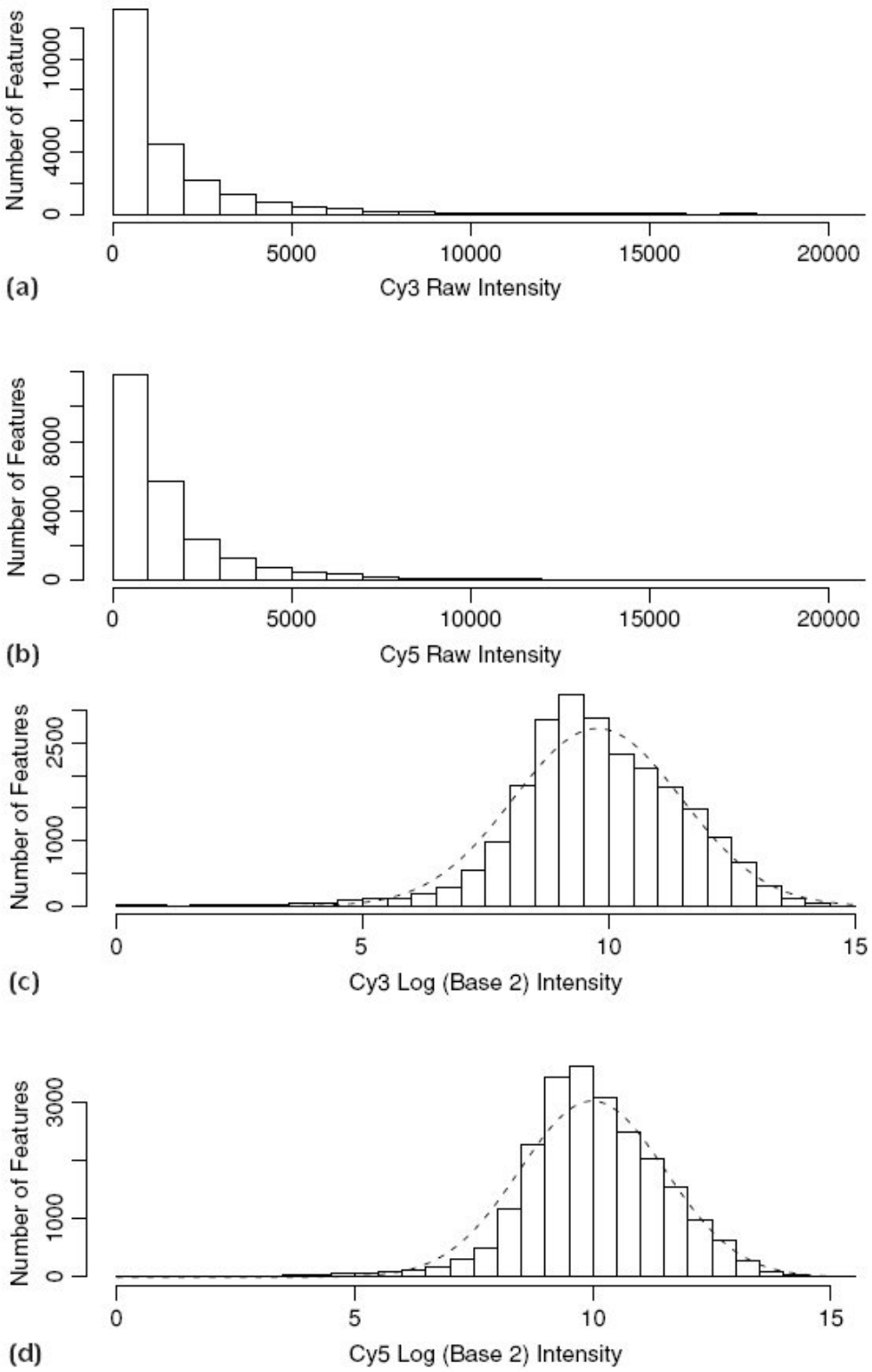


Figura 5.2: Istogramma delle intensità grezze e del logaritmo delle intensità di Cy3 e Cy5. Gli istogrammi delle intensità degli spot per i dati del fibroblasto umano. **(a)** Le intensità grezze per il canale Cy3; i dati sono deviati a destra con la maggior parte degli spots aventi bassa intensità, ed un numero decrescente di spots aventi alta intensità. **(b)** le intensità grezze per il canale Cy5; la struttura è la stessa come in (a). **(c)** Le intensità logaritmiche per il canale Cy3; le intensità assomigliano molto ad una curva normale con forma a campana (mostrata con linea tratteggiata). Vi è ancora un piccolo spostamento a destra, ma i dati logaritmici sono migliori per l'analisi rispetto ai dati sorgenti. **(d)** Il logaritmo delle intensità per il canale Cy5, insieme ad una curva normale (linea tratteggiata). Come con (c), le intensità sono approssimativamente normali, con piccolo spostamento a destra.

5.3 Normalizzazione interna all'array

L'insieme dei dati 5A è un esempio di una classe piuttosto tipica di esperimenti con microarrays. Gli sperimentatori si accingono ad usare il microarray per confrontare due differenti campioni ed identificare i geni che sono espressi in modo differenziale. In pratica, i due campioni vengono marcati con differenti fluorocromi in due reazioni chimiche separate, e le loro intensità vengono misurate con due differenti laser operanti a due differenti lunghezze d'onda. Inoltre, gli spots sull'array sono distribuiti su regioni diverse della superficie dell'array. Quando misuriamo differenti espressioni tra i due campioni, è necessario assicurarsi che le misure rappresentino la vera espressione differenziale del gene, e non polarizzazione ed errori introdotti dal metodo sperimentale. È necessario essere in grado di confrontare le intensità di Cy3 e Cy5 sulla stessa base di partenza - ciò si ottiene eliminando le quattro sorgenti che sono la causa di errori sistematici:

- I fluorocromi Cy3 e Cy5 possono essere incorporati in modo differenziale in DNA di differente abbondanza.
- I fluorocromi Cy3 e Cy5 possono avere differenti risposte di emissione (all'eccitazione laser) a differenti abbondanze.
- Le emissioni Cy3 e Cy5 sono misurate in modo differenziale; la misura può essere alterata dalla non linearità di risposta del tubo fotomoltiplicatore a differenti intensità.
- Le intensità di Cy3 e Cy5 misurate in differenti aree sull'array possono essere differenti poiché l'array non è orizzontale e pertanto la focalizzazione è differente nelle diverse regioni dell'array.

  Non è possibile distinguere le prime tre sorgenti di errori sistematici, e quindi queste risultano –nostro malgrado- combinate insieme. In questo paragrafo descriveremo tre metodi di correzione per differenti risposte dei canali di fluorocromi Cy3 e Cy5:
- Regressione lineare del fluorocromo Cy5 contro il fluorocromo Cy3
- Regressione lineare del rapporto logaritmico contro l'intensità media
- Regressione non lineare (Loess) del rapporto logaritmico contro l'intensità media

La polarizzazione spaziale puó essere corretta separatamente; descriveremo due metodi per correggerla:

- La regressione bidimensionale Loess
- La regressione Loess blocco a blocco

Tutti i metodi descritti in questo paragrafo si basano su una assunzione centrale**:** *La maggior parte dei geni sul microarray non sono espressi in modo differenziale.* Se questa assunzione è vera, allora questi metodi sono pregni di significato. Di converso, se questa assunzione non é vera, allora questi metodi possono non essere affidabili, e potrebbe essere appropriato un differente progetto sperimentale unitamente ad un diverso metodo di normalizzazione, come quello di usare un campione di riferimento.

Regressione Lineare del fluorocromo Cy5 contro Cy3

Il primo metodo - che è anche il piú semplice- per verificare se i canali dei fluorocromi Cy3 e Cy5 si stanno comportando in maniera confrontabile, si avvale del diagramma di dispersione delle intensità sui due canali (Figura 5.1b). Se i canali dei fuorocromi Cy3 e Cy5 si stanno comportando in modo simile, allora la nuvola dei punti sul diagramma di dispersione dovrebbe approssimare una linea retta, e la linea di regressione (che attraversa tutti i dati), dovrebbe avere una pendenza (ndt: coefficiente angolare) di 1 ed una intercetta di 0. Scostamenti da questi valori rappresentano risposte differenti dei canali di Cy3 e Cy5:

- Una intercetta diversa da 0 indica che uno dei canali è consistentemente più luminoso dell'altro.
- Una pendenza diversa da 1 indica che un canale risponde piú fortemente alle intensità rispetto all'altro.
- Deviazioni dalla linea retta indicano non linearità nella risposta della intensità sui due canali.

Esempio 5.2: Regressione lineare applicata all'insieme dei dati 5a

Viene disegnata una linea retta attraverso il diagramma di dispersione del logaritmo delle intensità dei dati provenienti dal fibroblasto umano, facenti parte dell'insieme di dati 5A (Figura 5.1b). L'intercetta della retta è 1.41 e la pendenza è 0.88. Questo significa che a basse intensità, il canale del fluorocromo C5y fornisce una risposta più elevata, mentre alle alte intensità è il canale del fluoro cromo Cy3 a fornire la risposta più elevata.. I punti più alti dei dati si incurvano discostandosi dalla linea retta. Ciò significa che la relazione tra i canali Cy3 e Cy5 non è completamente lineare. L'interpolazione con una linea retta può essere utilizzata per normalizzare i dati. La procedura è immediata:

1. Graficare Cy3 in funzione di Cy5 con un diagramma di dispersione.
2. Interpolare con una regressione lineare il diagramma di dispersione ed identificare l'intercetta ed il gradiente.
3. Sostituire i valori di Cy3 con i valori interpolati sulla linea di regressione.

Questo metodo per la normalizzazione funziona bene per i dati per i quali l'interpolazione lineare è buona, e costituisce un ragionevole metodo preliminare per visualizzare i dati. Tuttavia, ci sono due svantaggi di questo metodo:

- L'occhio ed il cervello umano funzionano meglio nel percepire le differenze da linee verticali od orizzontali, piuttosto che da linee oblique. Pertanto, non è sempre facile vedere la non linearità dei dati con questo tipo di grafico.
- La regressione lineare tratta i canali dei fluorocromi Cy3 e Cy5 in modo differente, e ciò produrrebbe un risultato diverso se fosse rappresentato Cy3 contro Cy5.

Regressione Lineare del Rapporto Logaritmico contro l'intensità Media

Un metodo alternativo, ed invero molto utile, per visualizzare e normalizzare i dati è quello di produrre un diagramma di dispersione del rapporto logaritmico contro l'intensità media di ciascuno spot. Questi diagrammi sono, talvolta, denominati diagrammi MA nella letteratura dei microarray. In questi diagrammi, ciascun punto rappresenta uno spot, in cui la coordinata x assume il valore medio del logaritmo delle intensità di Cy3 e Cy5, e la coordinata y assume come valore la differenza tra il logaritmo delle intensità dei canali Cy3 e Cy5 (i.e., il rapporto logaritmico). Il diagramma MA viene denominato in questo modo poiché l'intensità media è talvolta denominata A, ed il rapporto logaritmico è talvolta denominato M. Il diagramma MA è correlato al diagramma di dispersione del logaritmo delle intensità dei due canali; esso può essere ottenuto dal diagramma di dispersione ruotandolo di 45 gradi, e quindi scalando in modo appropriato i due assi. Tuttavia, il diagramma MA è uno strumento più potente per visualizzare e quantificare entrambe le risposte differenziali dei canali Cy3 e Cy5 (sia quelle lineari che quelle non lineari). Per prima cosa appare, di solito, più chiaramente se i due canali stanno rispondendo in modo differenziale oppure in modo non lineare. Se i due canali si stanno comportando in modo simile, allora i dati dovrebbero apparire distribuiti in modo simmetrico intorno ad una linea

orizzontale che passi per lo zero; ogni differenza da questa linea orizzontale indica una differente risposta dei due canali. L'occhio ed il cervello umano funzionano meglio nel'elaborare linee orizzontali piuttosto che linee diagonali, pertanto è più facile visualizzare la differenza tra i due canali utilizzando i diagrammi MA piuttosto che i diagrammi di dispersione. Seconda cosa, qualsiasi regressione lineare e non, sviluppata sul logaritmo del rapporto contro l'intensità media, tratta i due canali allo stesso modo.

Quindi, tali regressioni risultano essere molto robuste e riproducibili, piuttosto che lo sviluppo di regressioni su un canale contro l'altro.

Esempio 5.3: Diagramma del logaritmo del rapporto contro l'intensità media per l'insieme dei dati 5A

Il logaritmo del rapporto per ciascuno spot è rappresentato nel diagramma contro la media dei logaritmi della intensità per tutti gli spots dell'insieme di dati 5A (Figura 5.3). I dati non sono simmetrici intorno alla linea orizzontale passante per lo zero: il canale del fluorocromo Cy5 risponde più fortemente alle basse intensità, mentre il canale Cy3 risponde più fortemente alle alte intensità. (Questa è la stessa conclusione a cui andiamo incontro usando la regressione diretta). Poiché assumiamo che la maggior parte dei geni non siano differenzialmente espressi, attribuiamo questo effetto ad un artefatto sperimentale e normalizziamo i dati per rimuovere questo effetto prima di verificare i geni espressi in modo differenziale. La normalizzazione lineare che usa il logaritmo del rapporto e l'intensità media si comporta in modo simile alla regressione lineare sui dati Cy5 e Cy3.

Vi sono quattro passi:

1. Costruire la media del logaritmo dell'intensità ed il rapporto logaritmico per ciascuno spot.
2. Produrre il diagramma MA.
3. Sviluppare una regressione lineare del rapporto logaritmico sulla media delle intensità logaritmiche.
4. Per ciascuno spot, calcolare il rapporto logaritmico normalizzato sottraendo il valore interpolato sulla regressione dal logaritmo del rapporto dei dati grezzi.

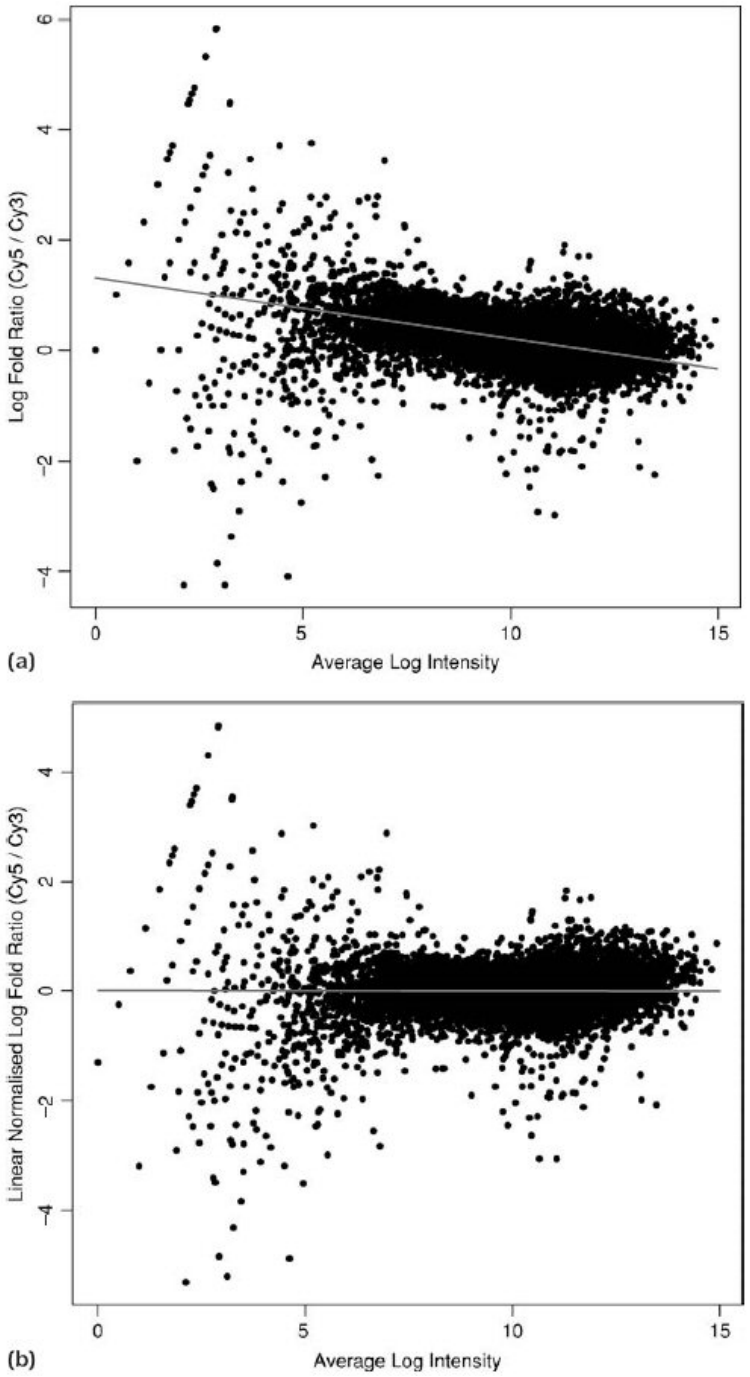


Figura 5.3: Diagrammi del logaritmo del rapporto in funzione dell'intensità media e normalizzazione lineare. Diagramma di dispersione del logaritmo del rapporto degli spots in funzione dell'intensità media per i dati dell'insieme 5A. Ciascun punto sul grafico rappresenta un differente spot. La coordinata x è l'intensità media dei canali Cy3 e Cy5; la coordinata y è il logaritmo del rapporto di Cy5 diviso Cy3 (uguale alla differenza nel logaritmo delle intensità tra Cy5 e Cy3). Questi diagrammi mostrano la tendenza media del logaritmo del rapporto come funzione dell'intensità. Questi diagrammi sono talvolta denominati "diagrammi MA "; essi sono geometricamente correlati al diagramma di dispersione di Cy5 in funzione di Cy3 (Figura 5.1) ottenuto ruotando il grafico di 45 gradi, e quindi scalando i due assi. **(a)** Una approssimazione lineare è stata operata per tutti i punti dell'insieme, che dimostra una chiara tendenza nelle risposte di Cy5 e Cy3. Alle basse intensità, il canale Cy5 risponde piú fortemente, mentre alle basse intensità è il canale Cy3 a rispondere più fortemente. Noi assumiamo che la maggior parte dei geni non siano differenzialmente espressi, così che questa linea rappresenta un artefatto sperimentale piuttosto che un'espressione differenziale. Il logaritmo del rapporto può essere normalizzato linearmente sottraendo il valore interpolato sulla linea retta da ciascun rapporto logaritmico. Benché la linea retta interpoli molto bene, non è corretto interpolare il centro dei dati. Alle alte intensità, i dati appaiono appiattirsi, suggerendo che una interpolazione non lineare potrebbe fornire risultati più affidabili. **(b)** I dati sono stati normalizzati con una linea di

regressione in (a) sostituendo il valore interpolato sulla linea da ciascun rapporto logaritmico per ciascuno spot. La linea di regressione è trasformata in una linea orizzontale attraverso lo zero. I punti di intensità più alti giacciono al di sopra della linea.

Esempio 5.4: Normalizzazione lineare sui dati del fibroblasto

La regressione lineare è applicata al diagramma MA sull'insieme dei dati 5A relativi ai fibroblasti (Figura 5.3a). La linea retta ha una intercetta di 1.31 ed una pendenza di -0.11; alle basse intensità, il canale Cy5 è più luminoso rispetto al canale Cy3, mentre alle alte intensità è il canale Cy3 ad essere più luminoso del canale Cy5. I valori interpolati con una linea retta sono sottratti dal logaritmo dei rapporti che sono usati per identificare i geni espressi in modo differenziale.

Regressione non Lineare del Logaritmo del Rapporto contro l'intensità Media

Se si guarda con attenzione alla Figura 5.3a, si vede che la singola linea retta non interpola perfettamente la nuvola dei dati: alle alte intensità la linea appare essere molto bassa. È molto comune con i dati di un microarray che la relazione tra i canali Cy3 e Cy5 non sia lineare. Quando ciò accade, la regressione lineare puó non essere la migliore risposta e, in questi casi, una qualche forma di regressione non lineare puó essere molto piú adatta. Il metodo piú comunemente utilizzato di regressione non lineare con i dati dei microarrays, è denominato regressione Loess (talvolta chiamata regressione Lowess*). Loess sta per locally weighted polynomial regression*. Il principio di funzionamento di tal regressione consiste nello sviluppare un elevato numero di regressioni locali in finestre sovrapposte nella direzione delle ascisse (Figura 5.4a) e quindi congiungere insieme le regressioni per formare una curva smussata (dolce nella sua evoluzione) (Figura 5.4b). La regressione Loess è oggetto di statistica relativamente avanzata, ed è generalmente disponibile, appunto, in pacchetti di software di statistica avanzata, come ad esempio *R* o MathLab. Tuttavia, è abbastanza semplice usarla se si possiede una conoscenza di base del modo in cui *R* lavora ed, in aggiunta, ci sono anche due pacchetti scritti specificamente per *R* per l'analisi dei dati di microarray che usano Loess per sviluppare la normalizzazione[6]. La regressione Loess è stata anche implementata - ed è commercialmente disponibile - in molti software per l'analisi dei dati di microarrays.

La regressione Loess è sviluppata in due passi:

1. Costruzione della intensità logaritmica media e del rapporto logaritmico per ciascuno spot.
2. Produzione del diagramma MA.
3. Applicazione della regressione Loess ai dati.
4. Per ciascuno spot, calcolo del rapporto logaritmico normalizzato sottraendo il valore interpolato con la regressione Loess, dal rapporto logaritmico dei dati grezzi.

[6] Le URL per questi packages sono fornite alla fine del capitolo

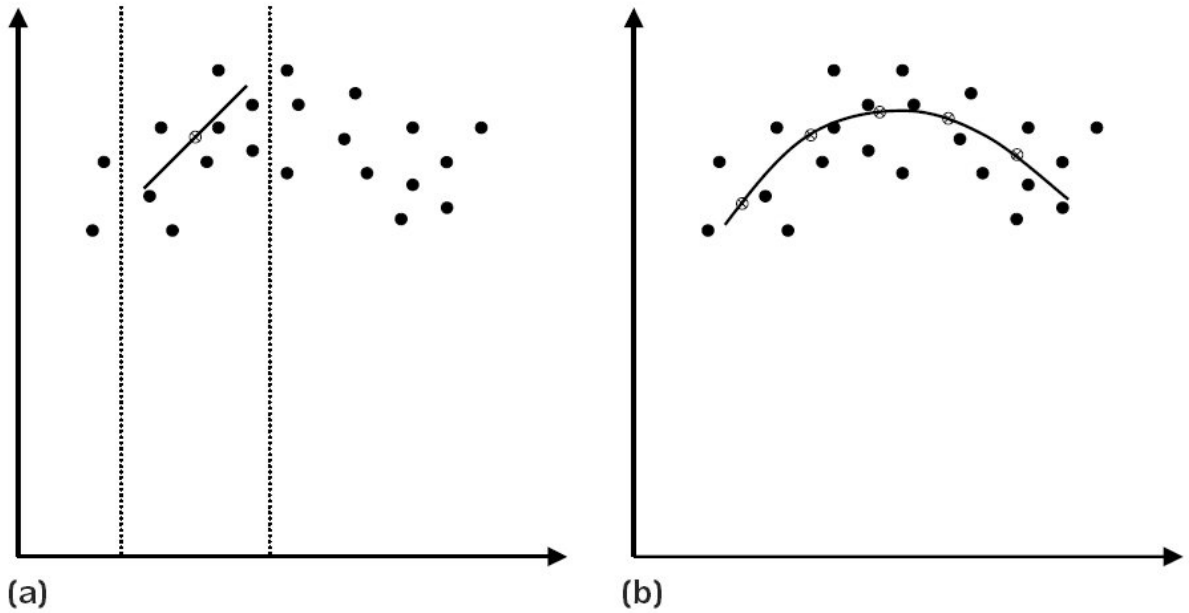


Figura 5.4: Normalizzazione Loess. **(a)** La regressione Loess lavora sviluppando un elevato numero di regressioni locali in finestre sovrapposte fino a coprire l'intero intervallo dell'insieme dei dati. La curva di regressione è, di solito, sia una linea retta che una curva quadratica. (la implementazione di default di R è una curva quadratica.) Ciascuna regressione consiste in un punto centrale e in una linea di regressione, oppure in una curva relativa a quel punto. **(b)** I punti e le curve risultanti dalle regressioni locali sono combinati per formare una curva smussata lungo l'intera lunghezza dell'insieme dei dati.

Esempio 5.5: Regressione non lineare applicata all'insieme dei dati 5A

La regressione Loess è applicata all'insieme di dati 5A dei fibroblasti umani (Figura 5.5a)[7]. La curva interpola i dati molto bene. I dati normalizzati (Figura 5.5b) sono bilanciati rispetto allo zero e sono pronti per l'analisi dell'espressione differenziale dei geni.

Benché Loess sia una tecnica statistica avanzata, è importante ricordare che essa non è niente di piú che un metodo computazionale della migliore curva interpolante attraverso una nuvola di punti. Non c'è nessun ragionamento teorico o concettuale relativo alla curva prodotta da Loess: essa è soltanto uno scaling di dati. La regressione Loess ha un certo numero di parametri che possono essere impostati dall'utilizzatore, ed i cui valori impatteranno sul modo in cui la curva stessa interpola i dati. Il più importante di questi è la grandezza della finestra, che determina il grado di "smussatezza" della regressione. Se la finestra è troppo piccola, la curva sarà molto sensibile alle salite ed alle discese locali dei dati, e sarà molto "contorta" (Figura 5.6a). Se la finestra è troppo larga, la curva sarà molto "rigida" e non sarà in grado di interpolare i dati in maniera efficace (Figura 5.6b).

[7] Il software statistico R é molto bene equipaggiato per sviluppare la regressione Loess. Essa puó essere trovata nel pacchetto *modreg.* Supponiamo che l'insieme dei dati si trovino in un insieme dati denominato *fibroblast,* con le variabili *average* e *lratio* contenenti la media dell'intensità logaritmiche ed il rapporto logaritmico. Quindi il comando R per ottenere la normalizzazione Loess dovrebbe essere:

```
attach(nonflat)
lmodel <- loess(lratio~x+y)
nonflat$lrationorm <- lratio - predict.loess(lmodel,nonflat)
```

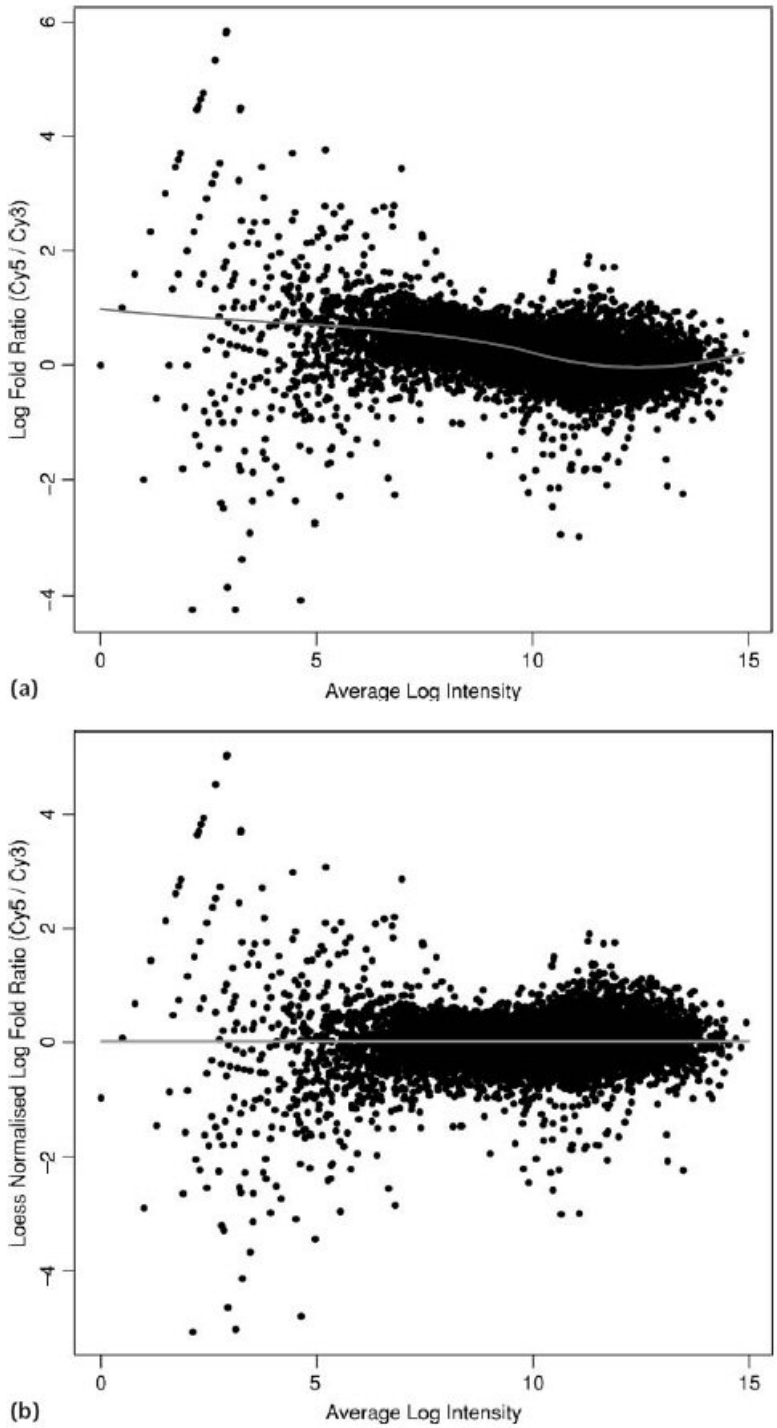


Figura 5.5: Normalizzazione Loess. **(a)** Il diagramma di dispersione dei dati del fibroblasto è rappresentato con un grafico con una interpolazione Loess tra tutti i dati. La curva non lineare sembra interpolare la forma dei dati meglio di una regressione lineare. Il rapporto logaritmico può essere normalizzato in modo non lineare sottraendo il valore interpolato sulla curva Loess tra i dati. Questo metodo funziona bene per i dati in cui vi sia una relazione non lineare tra le risposte dei canali Cy3 e Cy5. **(b)** I dati del fibroblasto dopo la normalizzazione Loess. La linea orizzontale attraverso lo zero corrisponde alla linea curva in (a). La linea orizzontale sembra attraversare il centro della nuvola dei dati molto bene.

Correzione degli Effetti Spaziali

In alcuni esperimenti con i microarrays, vi è una polarizzazione spaziale dei due canali: in qualche regione dell'array il canale Cy3 è piú luminoso, mentre in altre regioni dell'array è il canale Cy5 ad essere più luminoso. Questo può dipendere dal fatto che l'array non è perfettamente piatto od orizzontale nello scanner. La profondità di fuoco dei due laser è diversa - la profondità di fuoco è proporzionale alla lunghezza d'onda, e quindi è maggiore in Cy5 che in Cy3. Se l'array non è orizzontale, è dunque possibile che in qualche regione dell'array i due laser siano a fuoco, mentre in altre regioni dell'array il canale Cy5 potrebbe essere a fuoco, ma il canale Cy3 potrebbe essere leggermente fuori fuoco. Questo fatto può riflettersi sul logaritmo dei rapporti, con alcune regioni dell'array che presentano, in generale, rapporti logaritmici positivi, ed altre regioni che presentano rapporti logaritmici negativi.

Quando questo accade, è possibile correggere la polarizzazione spaziale usando diverse tecniche di normalizzazione: la regressione Loess bidimensionale e la regressione Loess blocco a blocco.

Regressione Loess Bidimensionale

Questo è in generale il metodo migliore per correggere la polarizzazione spaziale di un array. La regressione Loess bidimensionale lavora in modo simile alla Loess monodimensionale, ma invece di interpolare una curva, essa interpola una superficie sui dati per mezzo di polinomi bidimensionali. Per sviluppare una regressione Loess bidimensionale sui dati di un microarray, si devono compiere i seguenti passi:

1. Calcolo del rapporto logaritmico per ciascuno spot sull'array.
2. Produzione di un diagramma a falsi colori dei rapporti logaritmici degli spots in funzione delle coordinate x e y degli spots sull'array.
3. Sviluppo di una interpolazione Loess bidimensionale dei rapporti logaritmici in funzione delle coordinate x e y sugli spots.
4. Calcolo, per ciascuno spot, del rapporto logaritmico normalizzato sottraendo il valore interpolato sulla superficie Loess dal rapporto logaritmico dei dati grezzi.

Esempio 5.6: Insieme dei dati 5b. Polarizzazione spaziale su un array rene-fegato

In un esperimento a microarray -per guardare alla differenza tra l'espressione del gene nel rene e nel fegato di topo- i campioni del rene e del fegato provenienti dallo stesso topo sono stati preparati con i fluorocromi Cy3 e Cy5 ed ibridizzati al microarray.[8] I dati provenienti dall'array mostrano Cy3 piú luminoso all'angolo superiore sinistro e Cy5 piú luminoso all'angolo inferiore destro (Figura5.7a). Per correggere questa polarizzazione, i dati sono stati interpolati con una superficie bidimensionale Loess (Figura 5.7b).[9]

[8] Questi dati sono stati ottenuti privatamente dalla Microarray Facility al Mammalian Genetics Unit nei laboratory del Medical Research Council Laboratories ad Harwell nello Oxfordshire, UK

[9] La regressione Loess bidimensionale puó essere sviluppata in R usando la stessa funzione Loess nel modreg package come se venisse usata per una regressioe Loess monodimensionale. Supponiamo che i dati rene-fegato siano contenuti in un insieme dati denominato nonflat, con le variabili x, y, lratio corrispondenti alla coordinata x, alla coordinata y ed al rapporto logaritmico di ciascuno spot. Quindi, per sviluppare la normalizzazione Loess bidimensionale, usiamo i seguenti comandi:

```
attach(nonflat)
lmodel <- loess(lratio  x+y)
nonflat$lrationorm <- lratio – predict.loess(lmodel,nonflat)
```

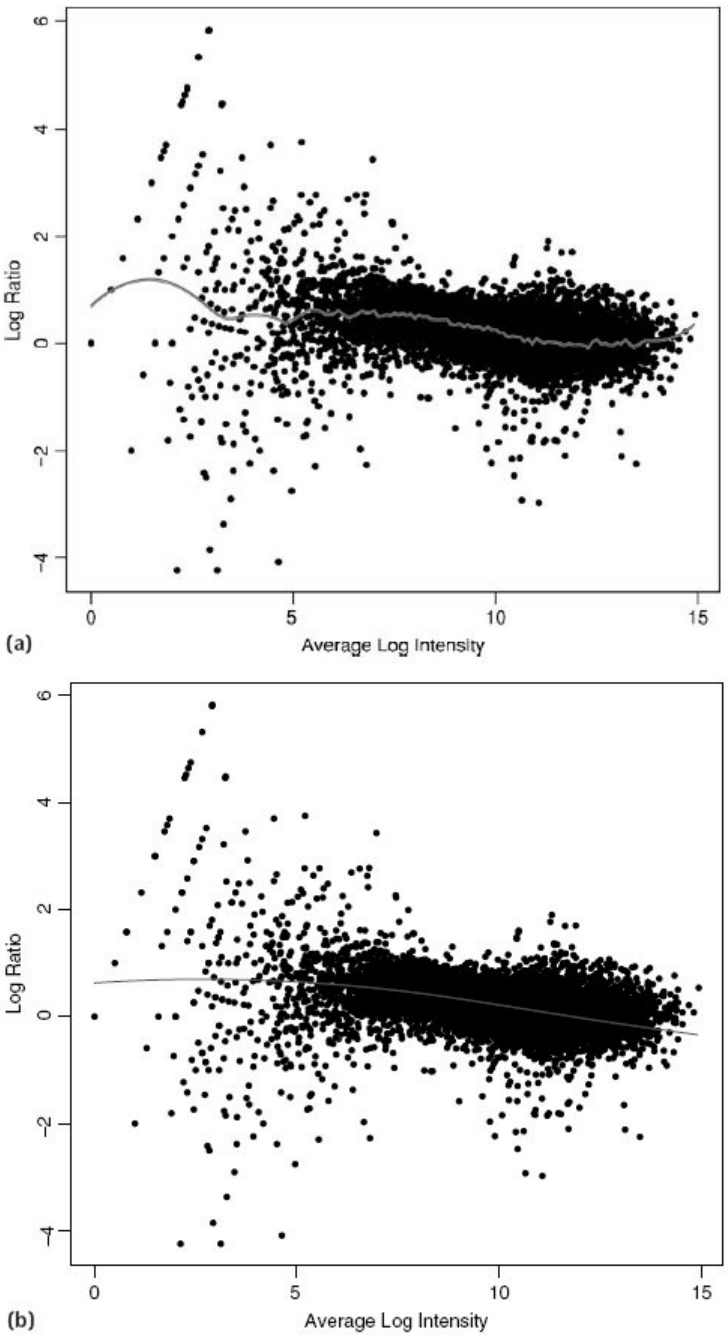
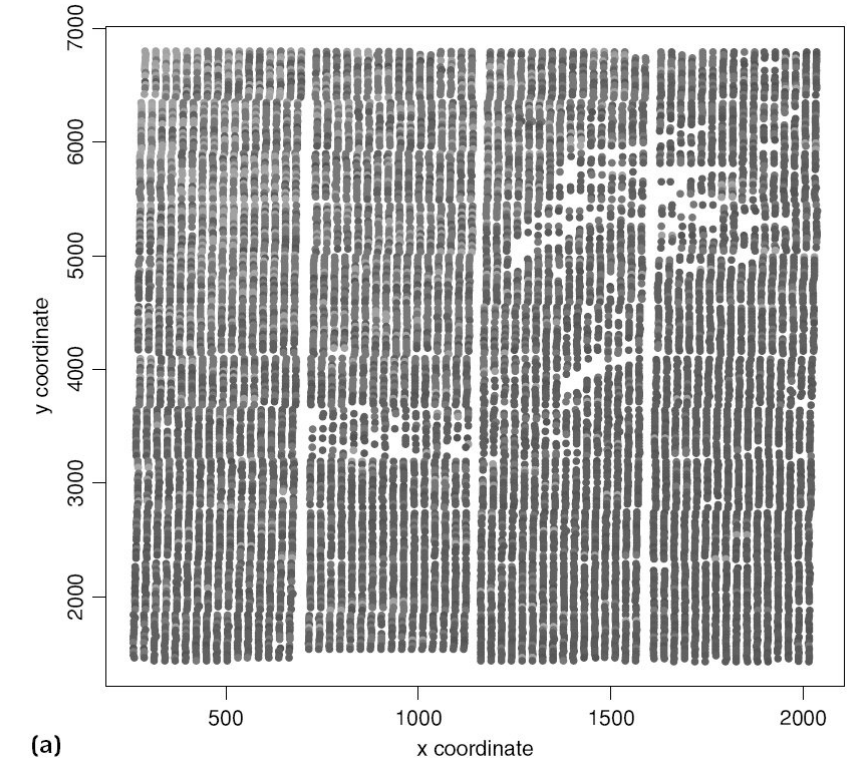


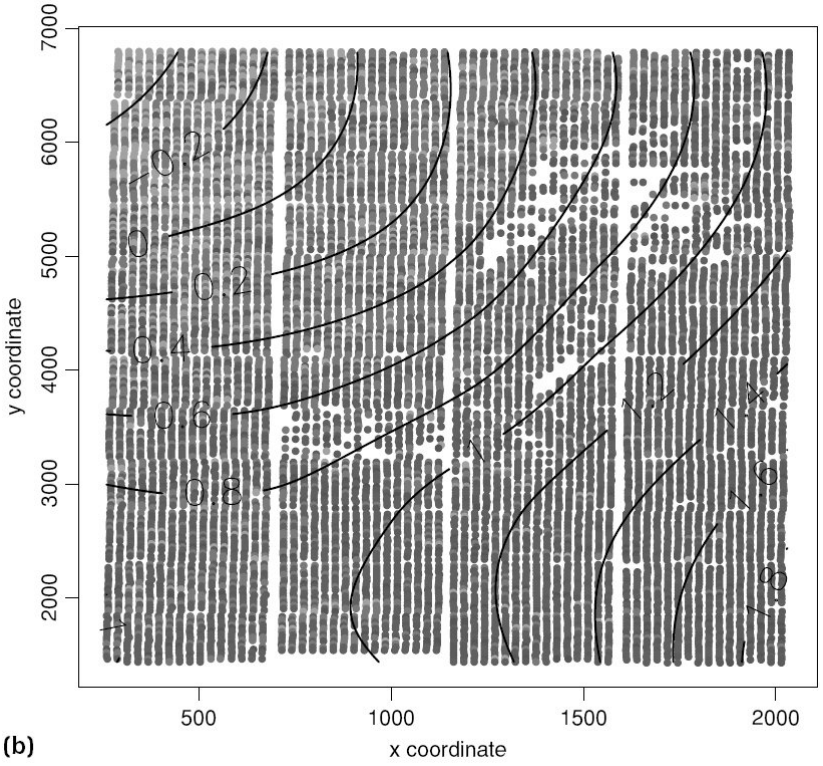
Figura 5.6: Robustezza della regressione Loess. La regressione Loess è applicata ai dati del fibroblasto usando finestre di diversa ampiezza. (a) La finestra per la regressione Loess è troppo piccola. La curva Loess segue gli spots locali troppo da vicino e come risultato abbiamo una curva molto rigida. (b) La finestra per la regressione Loess e troppo larga. La curva Loess è troppo instabile e non segue bene l'andamento dei dati.; la curva locale scende al di sotto del gruppo di geni fortemente espressi.

Dal diagramma è possibile vedere i contorni rappresentanti il gradiente differenziale delle intensità sugli arrays dall'alto a sinistra fino al basso a destra. I valori interpolati dalla

superficie Loess sono sottratti da ciascuno spot per produrre un insieme di dati normalizzati senza la polarizzazione spaziale (Figura 5.7c).



(a)



(b)

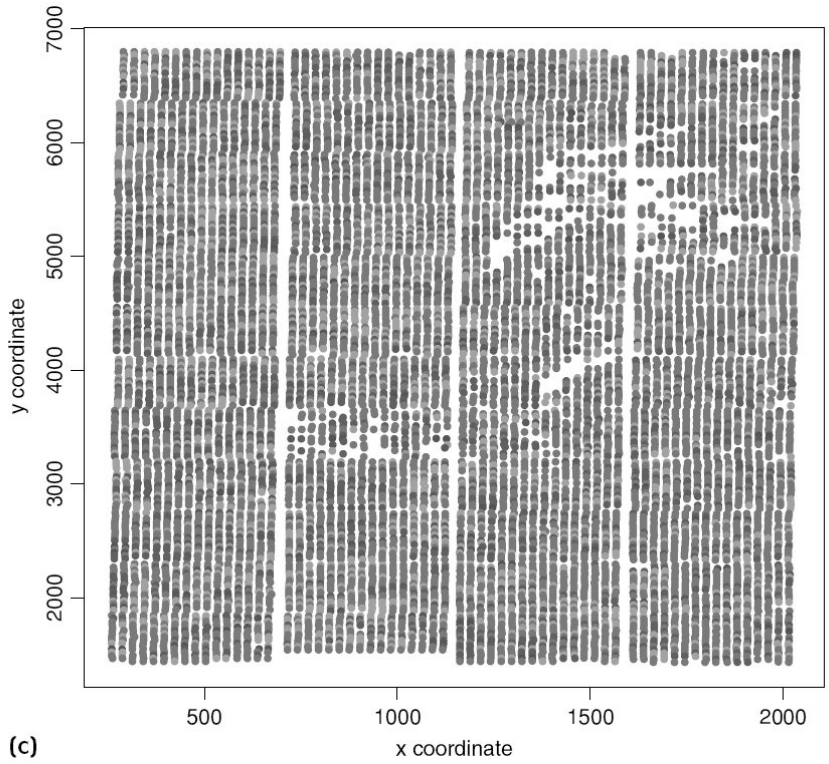


Figura 5.7: Polarizzazione spaziale su un microarray e regressione bidimensionale di Loess. (a) Rappresentazione a falsi colori del logaritmo dei rapporti di un microarray, con il rene del topolino in Cy3 ed il fegato dello stesso topolino in Cy5 (data set 5B). Ciascun spot rappresenta una caratteristica. Le coordinate x ed y di ciascun spot corrispondono alle coordinate x ed y della caratteristica nell'array. Il colore dello spot rappresenta il logaritmo del rapporto (Cy5/Cy3) della caratteristica, dove gli spot rossi hanno un logaritmo del rapporto positivo e gli spot verdi hanno un logaritmo del rapporto negativo. Vi è una marcata polarizzazione spaziale sull'array, dove gli spots verdi si trovano sull'angolo sinistro in alto e gli spots rossi nell'angolo destro in basso. Le aree dell'array in cui mancano gli spots rappresentano le caratteristiche che sono state marcate dal software di image processing, oppure quelle caratteristiche con background piú alto rispetto al segnale e che sono state rimosse dal set dei dati. **(b)** Gli stessi dati ma con una curva bidimensionale interpolante di Loess del logaritmo dei rapporti sovrapposti come zona di contorno. I contorni seguono il trend del colore, che vanno da un valore negativo alla sommità sinistra, ad un valore positivo in fondo a destra. **(c)** Diagramma a falsi colori di valori normalizzati del rapporto logaritmico delle caratteristiche. Questi sono calcolati sottraendo i valori interpolati della superficie di Loess dai valori grezzi dei rapporti logaritmici. Non vi è alcuna polarizzazione spaziale sui dati normalizzati.

Regressione Loess Blocco a Blocco

Un secondo metodo per correggere la polarizzazione spaziale sull'array consiste nello sviluppo della regressione Loess monodimensionale del rapporto logaritmico in funzione della intensità logaritmica media, ma invece di applicare questo metodo all'intero array (come veniva fatto in precedenza), il metodo viene applicato a ciascuna griglia sul'array separatamente. Questo metodo funziona bene quando viene introdotta un polarizzazione da differenti puntali sullo spotting robot, e le intensità dei i fluorocromi Cy3 e Cy5 si comportano in modo diverso per puntali diversi.

Esempio 5.7: Normalizzazione blocco a blocco sui dati dell'insieme 5b

La normalizzazione blocco a blocco è applicata all'insieme di dati 5B che appartengono all'esperimento rene-fegato. Vi sono 48 griglie sull'array (12 x 4). Il rapporto logaritmico è

normalizzato alla intensità media usando una regressione Loess separata per ciascuna griglia (Figure 5.8a e 5.8b). Dopo la normalizzazione, non vi è polarizzazione spaziale sull'array (Figura 5.8c). Vi sono due svantaggi con questo metodo. Primo, il numero di punti dei dati per ciascuna griglia, può potenzialmente essere abbastanza piccolo, e quindi è possibile che la maggioranza degli spots che stanno all'interno di un certo intervallo di intensità potrebbero essere espressi differenzialmente. Questo contravverrebbe al requisito che la maggior parte dei geni non siano differenzialmente espressi. La regressione Loess interpolerebbe i geni differenzialmente espressi, e così importanti informazioni sarebbero perdute durante il processo di normalizzazione.

Secondo, è piuttosto comune che la polarizzazione spaziale si instauri da un array non posto orizzontalmente nello scanner; fatto questo che può non avere alcuna relazione con la variabilità tra differenti puntali.

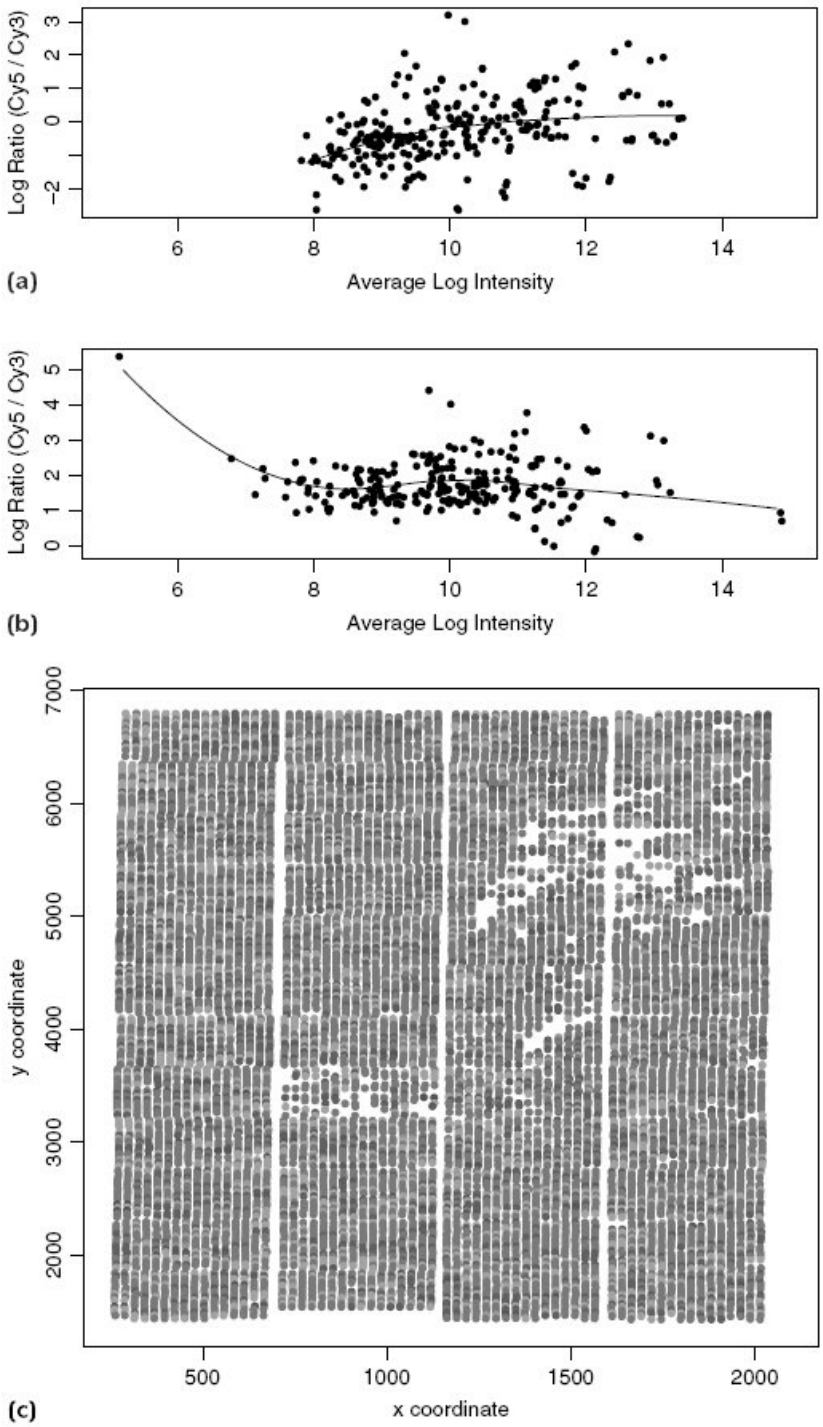


Figura 5.8: Regressione *Blocco a Blocco*. La regressione *blocco a blocco* vene sviluppata applicando la normalizzazione Loess monodimensionale agli spots in ciascuna griglia sugli arrays separatamente. L'array nell'insieme dei dati 5B ha 48 griglie. **(a)** Diagramma MA degli spots della griglia alta a sinistra; la maggior parte dei rapporti logaritmici sono negativi, ed in corrispondenza a questa regione l'array è di colore verde. L'interpolazione Loess appare essere buona, e quindi questi spots sono ben normalizzati. **(b)** Diagrammi MA per gli spots nella parte bassa a destra della griglia: la maggior parte dei rapporti logaritmici sono positivi. In corrispondenza a questa regione l'array è di colore rosso. Vi è un singolo spot che è debolmente espresso ma che presenta un rapporto logaritmico molto alto. Poiché non vi è un insieme di dati con abbastanza punti, la curva Loess ha interpolato questo punto in modo tale che esso appaia non essere differenzialmente espresso. Questo è un problema con la normalizzazione Loess blocco - a - blocco. **(c)** L'intero array è stato normalizzato usando la normalizzazione Loess blocco – a - blocco. La polarizzazione spaziale è stata eliminata.

5.4 Normalizzazione tra diversi Arrays

Nel Paragrafo 5.3 sono stati descritti i metodi di normalizzazione che possono essere usati per confrontare i canali dei fluorocromi Cy3 e Cy5 nel'ambito di un singolo array. In questo paragrafo tratteremo dei metodi di normalizzazione che ci permettono di fare confronti tra campioni ibridizzati a differenti arrays, che potrebbero essere sia arrays a due colori, che arrays Affimetrix. In tali esperimenti, ciascuna reazione di ibridizzazione puó essere leggermente differente, e cosí le intensità globali di differenti arrays possono essere differenti. Per essere in grado di confrontare (da una stessa base di partenza), i campioni ibridizzati a differenti arrays, è necessario correggere i risultati a causa della variabilitá introdotta usando array multipli.

Visualizzare i Dati: Diagrammi Box

Il Diagramma Box è un metodo per visualizzare contemporaneamente parecchie distribuzioni (di dati). Esso è un metodo eccellente per confrontare le distribuzioni delle intensità logaritmiche o i rapporti logaritmici di geni su parecchi microarrays.

Un Diagramma Box mostra una distribuzione come una segmento centrale compreso entro due linee orizzontali denominate "baffi". La linea che passa per il centro del segmento rappresenta la media della distribuzione. La scatola stessa rappresenta la deviazione standard della distribuzione, mentre le linee orizzontali che imbrigliano il segmento rappresentano i valori estremi della distribuzione[10].

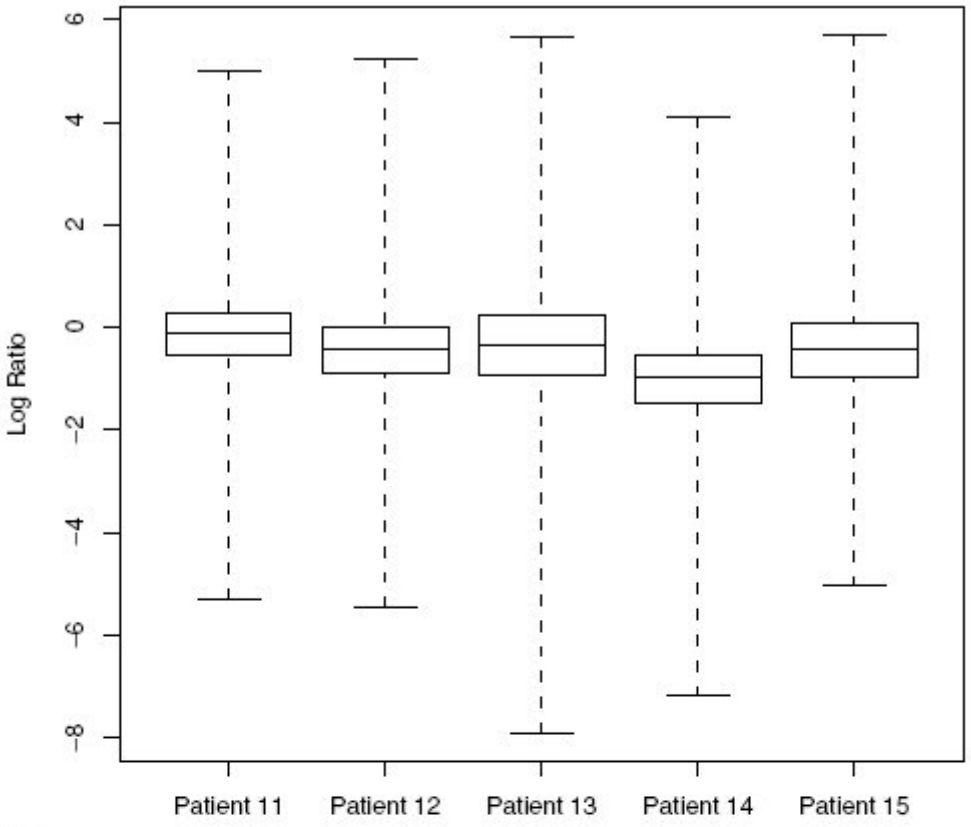
Esempio 5.8: Diagrammi box in pazienti con linfoma diffuso a grandi cellule di tipo B (insieme di dati 5C)

I campioni sono state presi da 39 pazienti sofferenti di un linfoma diffuso a grandi cellule di tipo B (DLBL) ed ibridizzati ai microarray, dove ciascun array contiene un campione nel canale Cy5 ed un campione di riferimento nel canale Cy3[11]. La Figura 5.9a mostra i diagrammi Box dei rapporti logaritmici dei campioni del paziente sui campioni di riferimento per 5 dei pazienti DLBL. Il diagramma Box ci permette di confrontare le distribuzioni del logaritmo dei rapporti in differenti pazienti. Per esempio, il paziente 14 ha un insieme più basso di rapporti logaritmici rispetto ad altri pazienti, ed il paziente 13 ha un range più ampio dei rapporti logaritmici rispetto ad altri pazienti. Vi sono tre metodi standard per la normalizzazione di dati simili ai dati dell'insieme 5C, in modo tale che gli arrays possano essere confrontati su una base comune. Essi sottostanno alla stessa assunzione fondamentale: le variazioni delle distribuzioni tra gli arrays sono il risultato di condizioni sperimentali e non rappresentano in alcun modo la variabilità biologica. Se questa assunzione non fosse vera, allora questi metodi non sarebbero appropriati. I tre metodi sono i seguenti:

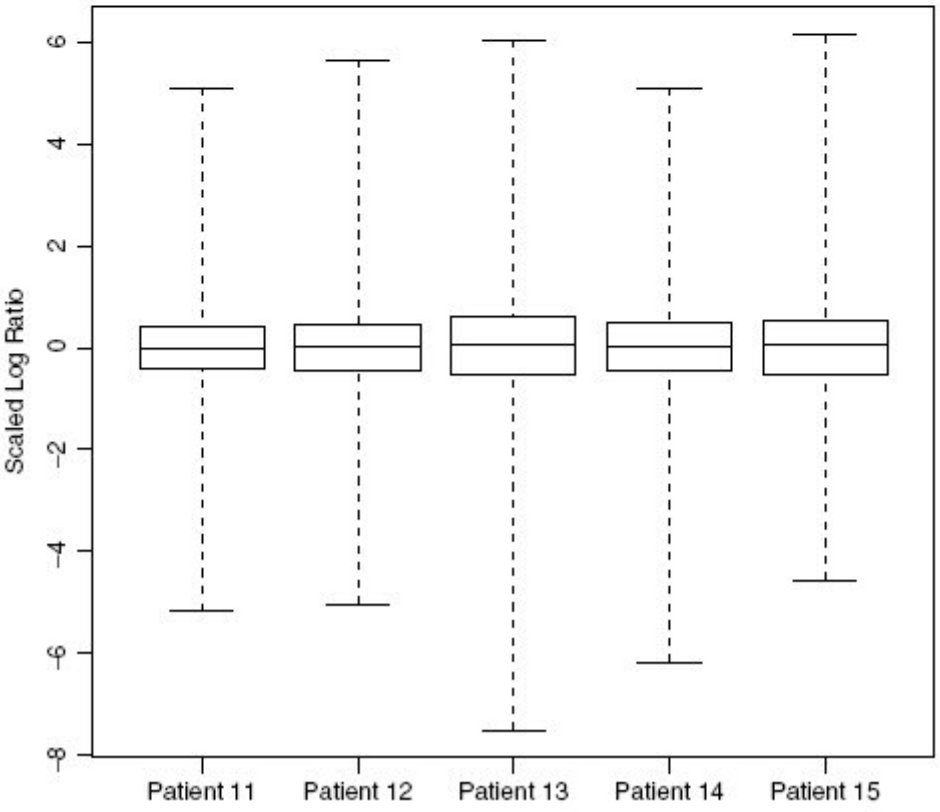
- Scaling
- Centraggio
- Normalizzazione della distribuzione

[10] Il diagramma Box che funziona in R é leggermente diverso e grafica la mediana della distribuzione al centro del Box, e la grandezza del Box rappresenta la deviazione del valore mediano assoluto dalla mediana. Questi sono robusti -non parametrici- equivalenti della media della mediana e della deviazione standard

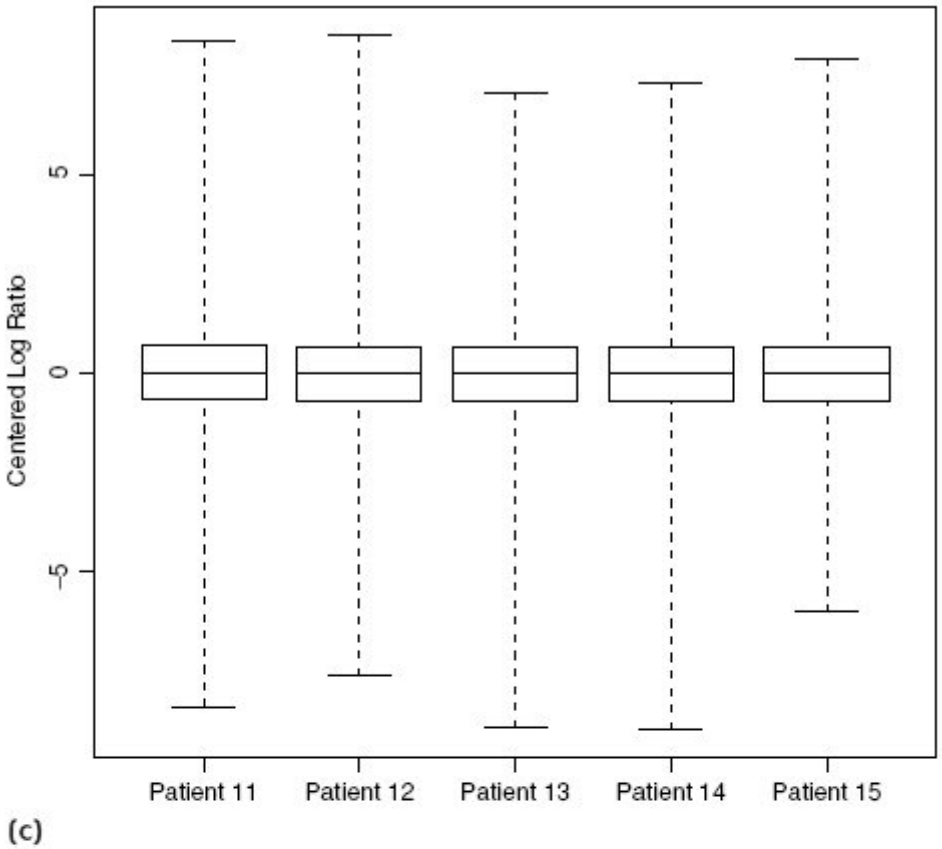
[11] Un riferimento della rivista da cui derivano questi dati é fornito alla fine del capitolo.


Log Ratio
6
4
2
0
−2
−4
−6
−8
Patient 11
Patient 12
Patient 13
Patient 14
Patient 15

(a)


Scaled Log Ratio
6
4
2
0
−2
−4
−6
−8
Patient 11
Patient 12
Patient 13
Patient 14
Patient 15

(b)



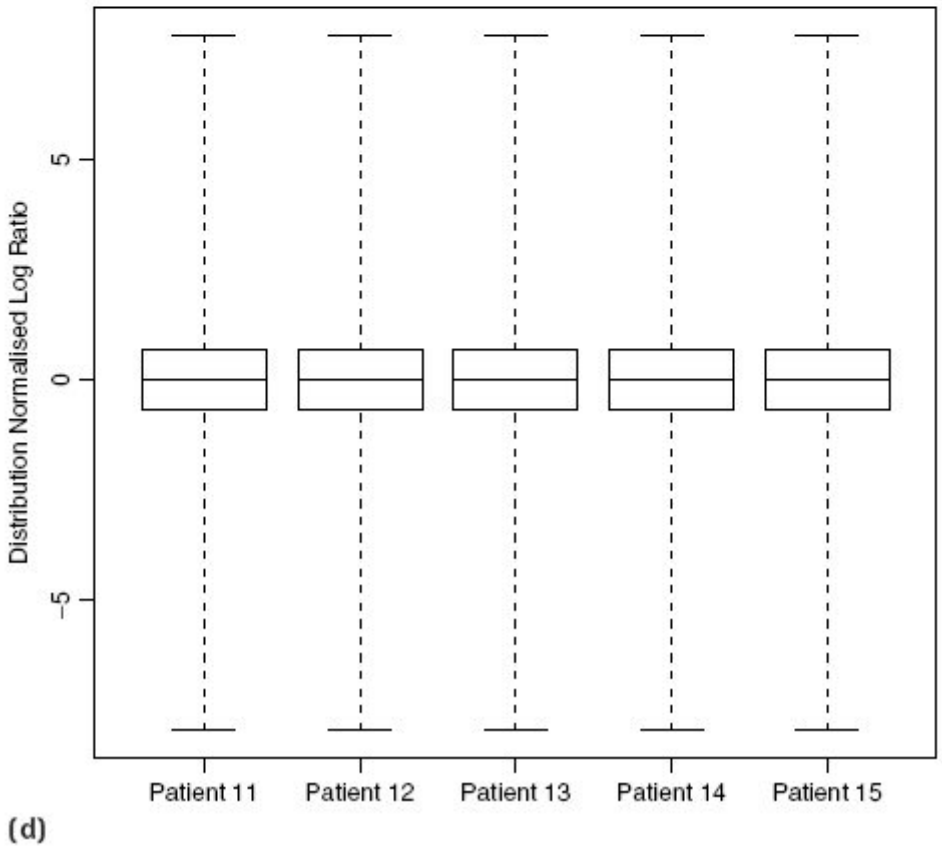


Figura 5.9: Scaling, Centraggio e normalizzazione della distribuzione. Differenti metodi che permettono il confronto tra i campioni su molti arrays per l'analisi dei dati come, ad esempio, le analisi di clustering (Capitolo 7) e l'analisi di classificazione (Capitolo 8). I dati in queste figure sono relativi a 5 pazienti

sofferenti di un linfoma diffuso a grandi cellule di tipo B (Insieme dei dati 5C). **(a)** Il diagramma box dei rapporti logaritmici dei dati grezzi relativi a 5 pazienti. La distribuzione dei rapporti logaritmici per tutti i pazienti è mostrata in un unico diagramma per modo che essi possano essere facilmente confrontati. La linea al centro di ciascun segmento rappresenta il valore della media (oppure il valore della mediana) della distribuzione. Le due linee orizzontali che imbrigliano il segmento (talvolta chiamate "baffi") rappresentano i valori estremi della distribuzione.
In questo diagramma i 5 pazienti presentano differenti valori della media, della deviazione standard e della distribuzione. **(b)** I dati sono stati scalati sottraendo la media della distribuzione da ciascun valore del rapporto logaritmico di ciascun paziente. Le medie delle distribuzioni sono uguali a zero. **(c)** I dati sono stati centrati sottraendo la media della distribuzione e dividendo per la deviazione standard. La distribuzione centrata per ciascun paziente ha la media uguale a 0 e la deviazione standard ugual ad 1. Il centraggio dei dati è utile quando si usi la correlazione come misura della distanza (Capitolo 7). **(d)** I dati sono stati normalizzati nella distribuzione in modo che ciascun paziente abbia lo stesso insieme di valori di misura; le distribuzioni per i cinque pazienti sono identiche.

Scaling

I dati sono scalati per assicurare che le medie di tutte le distribuzioni siano uguali (Figura 5.9b). Il metodo è semplice: si sottrae il rapporto logaritmico medio (oppure il logaritmo delle intensità) per tutti i dati di misura sull'array. La media delle misure su ciascun array sarà uguale a 0 dopo la normalizzazione. Un metodo alternativo all'uso della media è l'uso della mediana; ciò fornisce una misura più robusta della intensità media sul'array in situazioni dove ci siano "misure fuori posto" oppure le intensità non siano normalmente distribuite (non siano distribuite, cioè, secondo una curva normale).

Centraggio (dei dati)

I dati vengono centrati per essere sicuri che le medie e le deviazioni standard siano uguali per tutte le distribuzioni. (Figura 5.9c). Il metodo è simile allo scaling: a ciascuna misura sull'array viene sottratta la media delle misure sull'array e poi ogni misura viene divisa per la deviazione standard. Ancora, la media delle misure su ciascuno array sarà uguale a zero e la deviazione standard sarà uguale a1. Il centraggio è un metodo usato molto frequentemente per confrontare arrays multipli. Esso è particolarmente utile quando si calcoli il coefficiente di correlazione di Pearson di un grande numero di insiemi di dati, prima che venga effettuata l'analisi del cluster, poiché ci assicura che il coefficiente di correlazione definisca adeguatamente la misura di distanza dei dati. Questo aspetto è discusso approfonditamente nel capitolo 7.

Un'alternativa all'uso della media e della deviazione standard è l'uso della mediana e della deviazione assoluta della mediana (MAD). Questo ha il vantaggio di essere più robusto dell'uso della media e della deviazione standard, ma ha lo svantaggio di non produrre una misura di distanza adeguata quando si usa la correlazione di Pearson.

Normalizzazione della distribuzione

I dati della distribuzione sono normalizzati per essere sicuri che le distribuzione dei dati su ciascun array siano identiche. La metodologia è leggermente più complessa:

1. Centrare i Dati
2. Per ciascuno array, ordinare le misure centrate dalla più bassa alla più alta.
3. Calcolare una nuova distribuzione il cui valore piú basso sia la media dei valori del gene espresso piú debolmente su ciascuno arrays; ed il cui secondo valore

più basso sia la media del secondo valore più basso da ciascuno degli arrays; e così di seguito fino al valore più alto del valore medio dei valori piú alti da ciascuno degli arrays.

4. Sostituire ciascuna misura su ciascun array con la corrispondente media della nuova distribuzione. Per esempio, se una particolare misura è il 100esimo valore più grande sull'array, sostituirlo con il centesimo valore più grande nella nuova distribuzione.

Procedendo con la normalizzazione della distribuzione, le misure di ciascuno array avranno media uguale a zero, deviazione standard uguale ad 1, ed distribuzioni identiche per tutti gli arrays. La normalizzazione della distribuzione è una alternativa al centraggio come metodo per normalizzare i dati prima di sviluppare l'analisi del clustering (capitolo 7) o l'analisi della classificazione (capitolo 8). Essa è utile qualora ci si trovi con differenti arrays che abbiano differenti distribuzioni dei valori. Notiamo, in aggiunta, che il centraggio dei dati è semplice ed è il metodo utilizzato piú frequentemente per la normalizzazione dei microarray.

Esempio 5.9: Normalizzazione dell'insieme dei dati 5C

I Dati DLBL possono essere normalizzati usando tutti e tre i metodi. Lo scaling dei dati ci assicura che le medie dei rapporti logaritmici dei cinque pazienti siano tutti uguali a zero (Figura 5.9b). Tuttavia, le deviazioni standard sono alquanto differenti; per esempio, il paziente 13 ha una deviazione standard particolarmente grande. Il centraggio dei dati, in questo caso, assicura che le deviazioni standard siano tutte uguali ad 1 (figura 9.5c).

Tuttavia, le distribuzioni non sono identiche; per esempio, il paziente 14 ha qualche rapporto logaritmico particolarmente alto e negativo. La normalizzazione della distribuzione assicura che tutti gli arrays abbiano una identica distribuzione (Figura 5.9d).

Riassunto dei punti chiave

- La normalizzazione può rimuovere la variabilitá sistematica -indesiderata- per i dati del microarray.
- Visualizzare i dati con i diagrammi di dispersione e i diagrammi MA.
- Usare all'interno dell'array la normalizzazione per rimuovere l'effetto della polarizzazione dei fluorocromi e della polarizzazione spaziale.
- Usare la normalizzazione tra gli arrays per effettuare il confronto di arrays multipli.

Capitolo 6

Misura e quantificazione della variabilità dei microarrays

6.1 Introduzione

Nel Capitolo 5 sono stati descritti un certo numero di metodi per correggere la variabilità sistematica indesiderata sia nell'ambito dello stesso array che tra arrays differenti. In questo capitolo descriviamo i metodi di misura e quantificazione della variabilità aleatoria introdotta dagli esperimenti con microarrays. Le sorgenti più comuni della variabilità sono:

- La variabilità tra spots replicati sullo stesso array
- La variabilità tra due campioni marcati separatamente e ibridizzati allo stesso array
- La variabilità tra campioni ibridizzati ad arrays differenti
- La variabilità tra diversi individui di una popolazione, ibridizzati ad arrays differenti.

Stime di questa variabilitá sono essenziali per acquisire una comprensione di quanto correttamente si stia usando la piattaforma di microarray. Ci sono anche importanti parametri per determinare il numero di replicati necessari per un esperimento con microarray - un aspetto, questo, che verrà discusso approfonditamente nel Capitolo 10.

I primi due livelli di variabilità - tra gli spots replicati oppure tra campioni ibridizzati allo stesso array - sono significativi soltanto per arrays a due colori. Tuttavia, gli altri due livelli di variabilità - tra ibridizzazioni a differenti arrays e tra individui in una popolazione - sono significative sia per arrays a due colori che per arrays Affimetrix.

6.2 Misura e Quantificazione della Variabilità del microarray

Le variabilità tra differenti spots su un array, tra due campioni ibridizzati allo stesso array o tra campioni ibridizzati sono introdotte tutte dal processo sperimentale di realizzazione del microarray. Al contrario, la variazione tra individui in una popolazione è indipendente dal processo stesso del microarray. La variabilità sperimentale è misurata con la calibrazione degli esperimenti; la variabilità della popolazione è misurata con gli studi pilota.

Esperimenti di calibrazione

Lo scopo di un esperimento di calibrazione consiste nell' identificare e quantificare la sorgente di variabilità della piattaforma sperimentale di microarray di cui si dispone. L'informazione dovrebbe quindi essere usata per migliorare le procedure sperimentali, oppure per selezionare i livelli di replicazione tecnica per il controllo di qualità. È molto importante sviluppare un esperimento di calibrazione:

- Dopo aver allestito il laboratorio, e prima di intraprendere un qualsiasi progetto con i microarrays
- Dopo ogni cambio eventuale di un apparato di laboratorio, protocolli o persone; e/oppure
- Su base regolare per assicurare un livello di qualità continuato nel tempo

Un tipico esperimento di calibrazione include:

- Un progetto di array con parecchi spots replicati per ciascun gene; l'ideale sarebbe che questi spots fossero piazzati in locazioni diverse sull'array
- Produzione di un campione e marcatura con entrambi i fluorocromi Cy3 e Cy5; e/oppure
- Co-ibridizzazione del campione a diversi arrays.

In questo modo, sarà possibile misurare le variabilità introdotte da tutte e tre le sorgenti specifiche.

.

Studi Pilota

Lo scopo degli studi pilota è quello di fornire una guida approssimata del livello di variabilità in un popolazione prima di sviluppare un esperimento su larga scala.

Essi sono usati tipicamente prima di sviluppare una analisi intesa a calcolare il numero di replicati biologici necessari all'esperimento (Paragrafo 10.4).

Esempio 6.1: Studio Pilota

Cancro al seno in pazienti che sono trattati nel corso di 16 settimane con chemioterapia basata sul farmaco doxorubicina. I campioni saranno presi prima e dopo la chemioterapia ed ibridizzati ai microarrays. Noi vogliamo identificare i geni che sono espressi in modo differenziale come risultato della chemioterapia.

Nell'intenzione di dare un aiuto a progettare un esperimento su larga scala, sviluppiamo uno studio pilota per 5 pazienti per identificare il livello di variabilità della popolazione.

Quantificazione della variabilità

In statistica, è tipico rappresentare la variabilità di una popolazione con la deviazione standard. Negli esperimenti con microarray, l'espressione del gene (oppure la sua espressione relativa) su un particolare spot può essere immaginata come l'espressione vera del gene in un individuo o in un campione, a cui si aggiungono varie componenti di errore per ciascuna delle sorgenti di variabilità sperimentali (Figura 6.1). Ciascuna componente di variabilità può essere immaginata come una distribuzione separata con la propria deviazione standard. Per esempio, se noi consideriamo un array con parecchi spots replicati per ciascun gene, allora possiamo calcolare la deviazione standard della popolazione per le misure replicate dalla differenza tra le misure replicate di ciascun gene e la media dei replicati. La collezione di queste differenze per tutti gli geni può essere pensata come quella di un campione di variabile aleatoria rappresentante le differenze tra spots replicati. La deviazione standard dovrebbe dunque essere una stima della deviazione standard di questa variabile aleatoria e dovrebbe essere altresì una misura della variabilità tra i replicati. In questo esempio stiamo assumendo che la variabilitá tra spots replicati sia la stessa per tutti i geni. In seguito vedremo esempi dove queste assunzioni non sono valide.

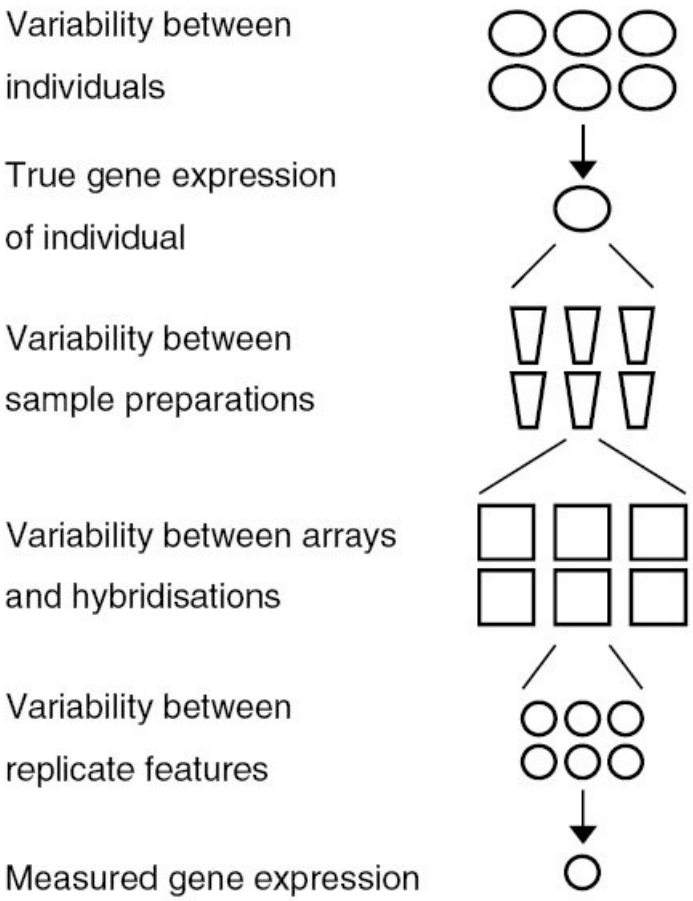


Figura 6.1: Sorgenti di variabilitá in un esperimento con microarray. Vi sono parecchi livelli di variabilitá nella misura dell'espressione del gene di uno spot. A livello più alto, c'è una variabilità biologica nella popolazione da cui i campioni provengono. Ad un livello sperimentale, c'è la variabilità tra le marcature e le preparazioni del campione, variabilità tra le ibridizzazioni dello stesso campione a differenti arrays, e la variabilità del segnale tra spots replicati sullo stesso array. La misura dell'espressione del gene include la vera espressione del gene, mescolata al contributo di ciascuna di queste variabilità

Distribuzione *log-normale*

Negli esperimenti con microarrays, è molto comune modellare la distribuzione degli errori di ciascun livello di variabilità esaminato, usando la distribuzione *log-normale*. Uno dei vantaggi derivanti dall'assumere la distribuzione log-normale, è che essa permette di esprimere i differenti livelli di variabilità come una percentuale, nota come coefficiente di variabilità. Il coefficiente di variabilità è uguale alla deviazione standard di un insieme di misure diviso per la media. Se la variabilità di un insieme di dati di microarray segue una distribuzione log-normale, allora la deviazione standard degli errori delle intensità sorgenti è proporzionale alla loro media (i.e., l'espressione sorgente del gene), e così il coefficiente di variabilità è ben definito. Nel modello log-normale, gli errori nei logaritmi delle intensità - e quindi anche gli errori dei rapporti logaritmici- seguono una distribuzione normale. Questo modello è corretto solo approssimativamente per i dati di un microarray reale. Noi vedremo nell'esempio che segue che la distribuzione degli errori tende ad avere un picco piú acuminato ed una *coda piú pesante* rispetto alla distribuzione normale. Questo significa che vi sono parecchi spots con errore più piccolo, ed un piccolo numero di spots con errori molto più grandi rispetto a quelli predetti dal modello log-normale. L'uso di una

distribuzione log-normale presuppone l'assunzione che l'ampiezza degli errori del logaritmo delle intensità è approssimativamente la stessa per spot di qualsiasi intensità.

Questo è approssimativamente vero per qualche insieme di dati del microarray, ma mostreremo anche parecchi esempi dove gli errori sono più grandi per spot di piccola intensità rispetto agli spot ad alta intensità. In tali casi, è possibile risolvere questo problema partizionando i dati in spots di bassa-intensità ed alta-intensità, e fornire piú di una misura di variabilità degli spots per differenti *ranges* di espressione. Se noi usiamo il modello log-normale, allora il coefficiente di variabilità (*v*) correla la deviazione standard degli errori distribuiti in modo normale nelle misure registrate, (*sigma),* attraverso la formula:

$$v = \sqrt{(\exp(\sigma^2) - 1)}$$

Equazione 6.1

Questa equazione dipende dall'uso dei logaritmi naturali (in base *e*); quindi se si stanno usando logaritmi in base 2, la deviazione standard deve essere moltiplicata per il ln(2) (approssimativamente 0.69) per ottenere un valore coretto di *sigma* per l'equazione 6.1

Metodo per la Misura della Variabilità

Abbiamo discusso quattro livelli di variabilità: tra gli spots, tra le ibridizzazioni, tra gli arrays e tra gli individui. Descriveremo ora un metodo per calcolare le deviazioni standard di ciascuno di questi livelli di variabilità. Ci saranno delle considerazioni leggermente diverse per ciascun caso; il metodo generale è il seguente:

1. Per ciascun insieme di replicati (spots, ibridizzazioni o individui), calcolare la media dei replicati.
2. Per ciascun replicato, calcolare la deviazione dalla media mediante il calcolo della differenza tra l'intensità del replicato e la media dell'insieme dei replicati.
3. Produrre i diagrammi MA delle deviazioni contro la media e verificare che la variabilità non sia dipendente dall'intensità.
4. Se lo si desidera, si può applicare alla deviazione una normalizzazione lineare o non lineare (vedere paragrafo 5.3).
5. Calcolare la deviazione standard della distribuzione degli errori usando tutti i replicati[1]. Se la variabilità dipende dall'intensità, potrebbe essere utile partizionare i dati in differenti *range* di intensità e calcolare la deviazione standard per ciascuna partizione.
6. Se l'assunzione log-normale è corretta, allora le deviazioni dovrebbero essere distribuiti in accordo ad una distribuzione normale. Questo può essere verificato con l'ausilio di un istogramma delle deviazioni.
7. Convertire la deviazione standard in coefficiente percentuale di variabilità moltiplicando per ln(2) ed applicare l'equazione 6.1.

[1] Benché l'intero insieme dei deviati replicati non siano variabili aleatorie indipendenti, la deviazione standard dei campioni é ancora uno stimatore non polarizzato della popolazione della deviazione standard e quindi questo procedimento é statisticamente significativo. Tuttavia, esso sarebbe non significativo per sviluppare test statistici per la normalità della distribuzione di tutte le deviazioni delle repliche.

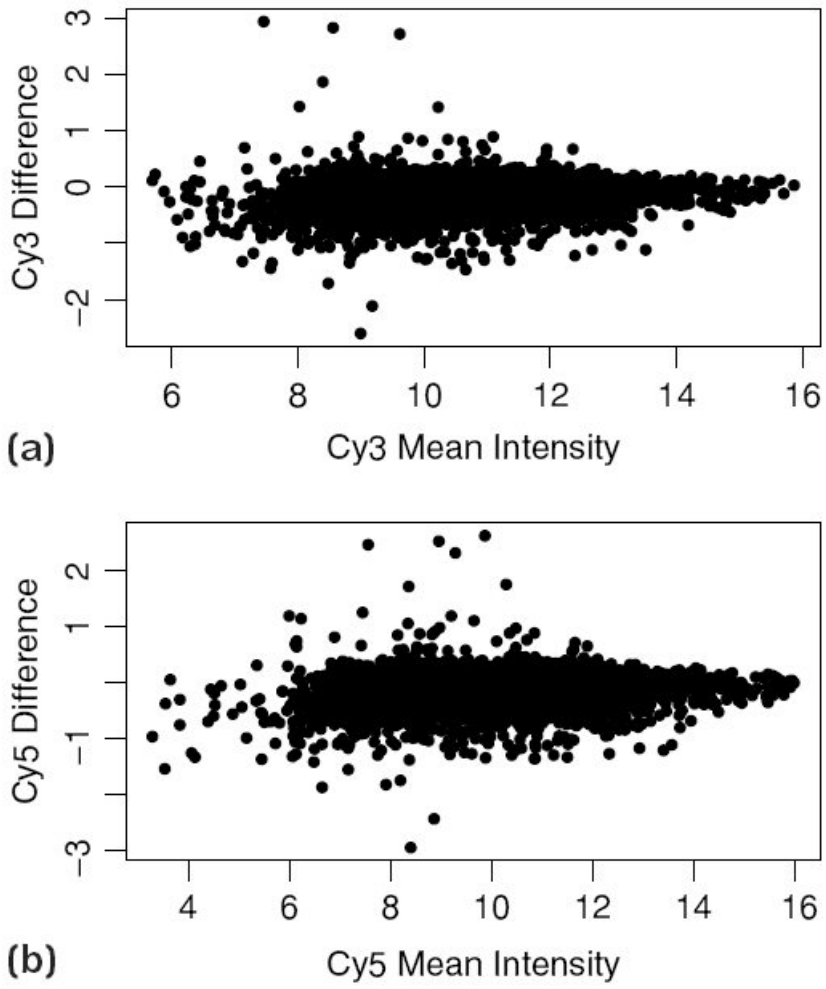


Figura 6.2: Diagramma a Punti della Variabilitá di Duplicati. Il diagramma a punti che mostra la relazione tra gli spots replicati su un array con l'insieme di dati 6A. Gli arrays sono stati costruiti con 6.000 geni spottati in duplicato. Un campione di rene di topo fu preparato e marcato due volte, una volta con il fluorocromo Cy3 ed un'altra con Cy5, ed entrambi i campioni furono ibridizzati all'array. Questo permette di stimare la variabilità tra spots duplicati, e tra due campioni marcati. Ciascun punto sulla figura rappresenta una coppia di geni duplicati per geni differenti. La coordinata x è la media del logaritmo (in base 2) delle intensità degli spots duplicati La coordinata y è la differenza tra l'intensità logaritmica del primo replicato e la intensità media. (a) I duplicati dal canale Cy3 sull'array. Vi sono due fenomeni da osservare. Primo, a tutte le intensità, la nuvola dei punti è generalmente negativa, implicando che il primo replicato è meno intenso del secondo replicato dello stesso gene. Vi sono tre ragioni perché questo fenomeno potrebbe essere ascrivibile a: (i) polarizzazione spaziale sull'array, (II) se i replicati sono depositati con differenti micro puntali, potrebbe esserci variabilità da micro puntali e a micro puntale; (III) se i replicati sono spruzzati con differenti micro puntali, più fluido potrebbe essere rilasciato dal micro puntale la seconda volta che esso è applicato sul vetrino. Questi effetti possono essere normalizzati usando la regressione Loess (Paragrafo 5.3). Secondo, benché la variabilità dei duplicati sia relativamente costante, essa decrementa al crescere dell'intensità. Potrebbe essere significativo produrre un report di due diversi coefficienti di variabilità tra i duplicati, uno per i geni debolmente espressi [log(2) intensità inferiore a 13], ed uno per i geni altamente espressi [log(2) intensità maggiore di 13]. (b) Un diagramma similare ma per il canale Cy5 sull'array. Questo diagramma è molto simile ad (a), ma con variabilità leggermente più grande.

Variazione tra gli spots replicati sull'array

La variazione tra spots replicati sull'array può essere misurata in un qualsiasi esperimento in cui sono stati usati gli spots replicati.

Esempio 6.2: Calcolo della variabilità tra spots replicati usando una self-self-ibridizazione (insieme di dati 6A)

In un esperimento per determinare la qualità di un kit di sviluppo di microarrays, l'RNA è stato estratto dal rene di topo e marcato due volte, la prima volta con il fluorocromo Cy3 e la seconda volta con il fluorocromo Cy5. I due campioni marcati sono stati ibridizzati ad un array con 6.000 geni *"spotted"* sul vetrino in duplicato[2]. I ricercatori volevano calcolare la variabilità tra spots duplicati.

La variabilità tra gli spots duplicati è calcolata separatamente per i due canali. I diagrammi MA (Figura 6.2) mostrano che la variabilità è approssimativamente costante; la variabilità è, infatti, più piccola per i geni espressi più fortemente. In aggiunta, il diagramma non è centrato sullo zero; indice, questo, di una polarizzazione sistematica tra i due replicati; questo può essere corretto usando la normalizzazione Loess (Paragrafo 5.3). Le distribuzioni normalizzate degli errori sono approssimativamente -ma non esattamente- normali, con gli errori aventi un picco più ripido e code più larghe rispetto alla distribuzione normale (Figura 6.3). In questo esempio, potremmo calcolare le deviazioni standard ed i coefficienti di variabilità sia dall'intero insieme di dati, che dalle partizioni di dati che fanno parte dei geni debolmente espressi e dei geni fortemente espressi, e calcolare deviazioni standard separate per le due partizioni.

Quando si usano tutti i dati, la deviazione standard è 0.25 per il canale Cy3 e 0.27 per il canale Cy5. Queste possono essere convertite in coefficienti di variabilità moltiplicando ciascuna deviazione standard per ln(2) ed applicando l'equazione 6.1; esse corrispondono a coefficienti di variabilità di 17 e 19 %, rispettivamente. Se partizioniamo i dati in geni debolmente espressi con intensitá log(base 2) minore di 13, ed i geni fortemente espressi con intensitá log(base 2) maggiore di 13, allora i coefficienti di variabilitá per i geni debolmente espressi sono 18 e 19% nei canali Cy3 e Cy5, rispettivamente, e per i geni fortemente espressi sono 14 % in entrambi i canali.

[2] I dati sono stati ottenuti privatamente dalla Microarray Facility alla the Mammalian Genetics Unit dei laboratori del Medical Research Council ad Harwell nello Oxfordshire, UK

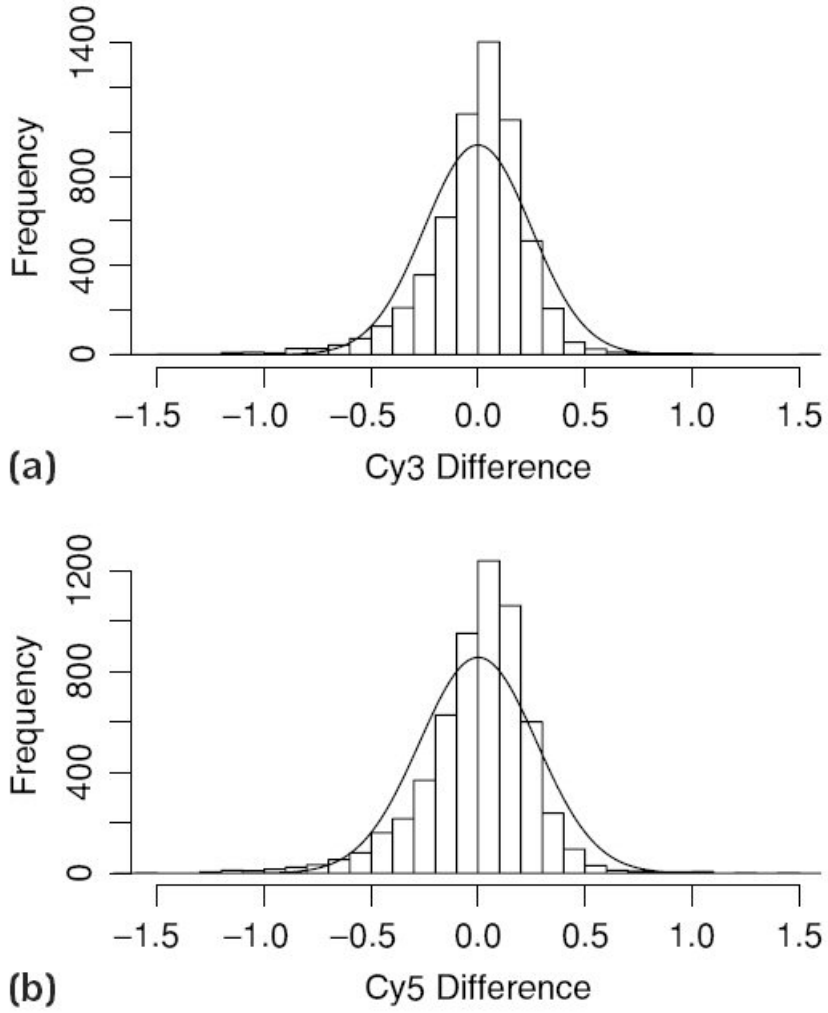


Figura 6.3: Distribuzione della Variabilità di Duplicati. La distribuzione della variabilità tra i duplicati derivante dall'insieme di dati 6A è mostrata con un istogramma, con la curva mostrante una distribuzione normale interpolata con la stessa media e deviazione standard. **(a)** Istogramma per il canale Cy3. La variabilità è a forma di campana. L'approssimazione normale è ragionevole ma non esatta: ci sono molti spots con variabilità più piccola che in una distribuzione normale e, corrispondentemente, vi sono molti spots fuori del limite, con variabilità più grande che in una distribuzione normale (questo è difficile vederlo in questo diagramma). È comune riferirsi a questo fenomeno come una distribuzione avente *code pesanti.* **(b)** Istogramma per il canale Cy5, mostrante le stesse proprietà come in (a)

Variabilitá tra i canali Cy3 e Cy5

La variabilità tra due campioni ibridizzati allo stesso array è misurata nel modo migliore con una self-self-ibridizzazione. Lo stesso campione biologico è marcato due volte, una volta con Cy3 ed un'altra volta con Cy5, e quindi misuriamo la variabilità tra i due insiemi di misura (dati). Questa variabilità è differente dalla polarizzazione sistematica tra i canali Cy3 e Cy5 discussa nel paragrafo 5.3. La polarizzazione sistematica rappresenta una consistente differenza introdotta dall'apparato sperimentale ed è rimossa usando i metodi di normalizzazione discussi al paragrafo 5.3. Ma, dopo che questa polarizzazione è stata rimossa, rimane la variabilità *random* tra i due insiemi di misura nei due canali. Questa è la variabilità che noi misuriamo.

Esempio 6.3: calcolo della variabilità tra i canali CY3 e cCY5

La self-self-ibridizzazione dei dati appartenenti all'insieme 6A, può essere usata per calcolare la variabilità tra due campioni marcati, ibridizzati all'array. La stessa procedura è applicata alla media delle misure degli spots duplicati in ciascun canale Cy3 e Cy5, che sono quindi normalizzati con la regressione Loess. In questo esempio la variabilità è approssimativamente costante ed approssimativamente distribuita come una curva normale. La deviazione standard e la distribuzione degli errori sono approssimativamente 0.18 (con il log in base 2), che corrisponde a un coefficiente di variabilità del 12%.

Variabilità tra Ibridizzazioni a differenti arrays

La variabilità di ibridizzazioni a differenti arrays consiste in due componenti: la variabilità relativa all'array stesso (in relazione alla loro produzione), e la variabilità relativa a differenti reazioni di ibridizzazione. Non è possibile misurare le due componenti separatamente (questo è conosciuto come effetto confusione delle variabili; ciò viene discusso approfonditamente nel Cap.10), e quindi le due variabili sono combinate in una singola misura. Per calcolare questa variabilità, ci sarà bisogno di ibridizzare gli stessi campioni marcati ad un certo numero di differenti arrays. Negli esperimenti dove é stato usato un campione di riferimento, il campione di riferimento su differenti arrays può servire a stimare questa variabilità.

Esempio 6.4: Calcolo della variabilità tra ibridizzazione di un campione di riferimento, ibridizzato ad arrays multipli (insieme di dati 6B)

In un esperimento relativo alla chemioterapia del cancro alla mammella, i campioni furono presi da 20 pazienti prima e dopo il trattamento con chemioterapia a base di doxorubicina[3]. I 20 campioni sono stati marcati prima del trattamento con Cy5 ed ibridizzati a 20 differenti arrays; ciascun array fu ibridizzato con lo stesso campione di riferimento che era stato marcato con Cy3. Vogliamo calcolare la variabilità tra le ibridizzazioni del campione di riferimento a differenti arrays.

Restringiamo l'analisi a 6350 geni in modo tale che tutti i dati siano presenti nel data set. Le intensità dei campioni di riferimento su ciascun array sono centrati (Paragrafo 5.3) per permettere un confronto tra gli arrays. Seguiremo la già menzionata procedura e produrremo il diagramma MA per tutti i dati (figura 6.4a). La deviazione standard è abbastanza costante a tutti i livelli di intensità, in modo che sia significativo parlare anche solo di un singolo coefficiente di variabilità. L'istogramma della distribuzione degli errori (figura 6.4b) mostra che la distribuzione è approssimativamente normale. La deviazione standard della distribuzione degli errori è 0.31. Moltiplichiamo questa per il ln(2) ed applichiamo l'equazione 6.1 per ottenere un coefficiente di variabilità del 22%. E dunque, in questo esperimento, vi è il 22% di variabilità tra gli stessi campioni marcati ibridizzati a differenti array.

[3] Un riferimento alla pubblicazione da cui quest'insieme di dati deriva é dato alla fine del capitolo

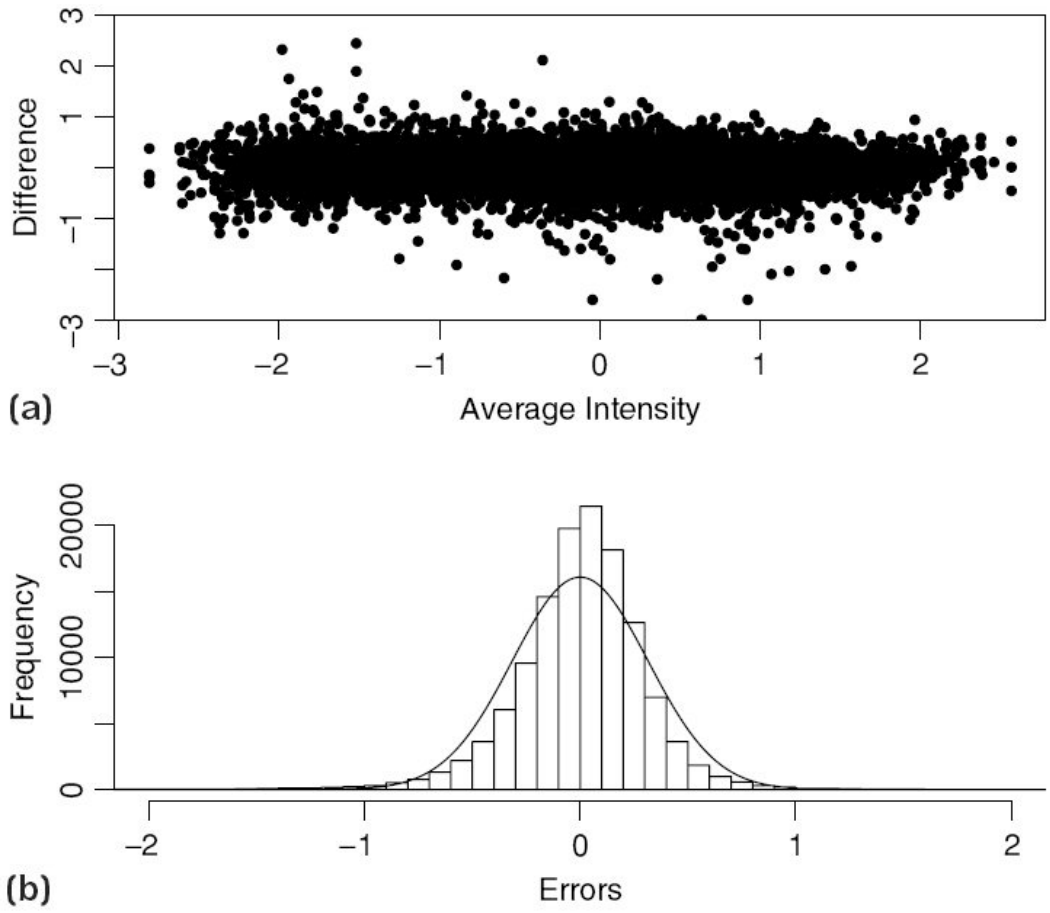


Figura 6.4: Variabilitá tra ibridizzazioni. Le figure visualizzano la variabilità tra ibridizzazioni dello stesso campione di riferimento a 20 differenti arrays nel'insieme dei dati 6B. Le intensità del campione su ciascun array sono state centrate così che esse abbiano media 0 e deviazione standard 1. **(a**) Diagramma a punti delle variabilità. A causa del'elevato numero di dati in questo esempio, il diagramma a punti mostra un campione statisticamente selezionato di replicati rappresentante il 10% dei dati originali. La variabilità è abbastanza costante sull'intero range di variabilità. **(b)** Istogramma di distribuzione dell'errore. La distribuzione è piú vicina ad una distribuzione normale rispetto al l'esempio dei spots duplicati, ma ci sono duplicati con variabilità più piccola rispetto a quella predetta dalla distribuzione normale, indicando anche che vi sono *code pesanti* di duplicati con variabilità estrema.

Variabilità tra individui

La variabilità tra individui in un popolazione è fondamentalmente differente dalle tre altre sorgenti di variabilità descritte precedentemente. Le altre sorgenti di variabilità sono tutte introdotte come parte del processo sperimentale. Non si tratta di variabilitá particolari, e ogni miglioramento nelle pratiche sperimentali che riducano quelle variabilità, sarebbe vantaggioso. La variabilità della popolazione, sull'altro versante, proviene dai sistemi biologici che si vogliono studiare. In molti esperimenti, noi siamo esplicitamente interessati alla variabilità tra gli individui, poiché potrebbe essere di una cera importanza sia nella malattia che nei risultati del trattamento. C'é da osservare, però, che la misura della variabilità tra individui è sviluppata nello stesso modo con cui vengono misurate le altre variabilità, e possono essere anche espresse come coefficiente di variabilità

Esempio 6.5: Calcolo della variabilità tra individui nell'insieme di dati 6B

Usando i dati dell'insieme 6B calcoliamo la variabilità dell'espressione del gene nella popolazione di pazienti con cancro alla mammella. Noi sviluppiamo la stesa procedura, ma invece di porre l'attenzione al campione di riferimento, poniamo l'attenzione al rapporto logaritmico di ciascun gene relativo al campione di riferimento come misura dell'espressione del gene. I rapporti logaritmici su ciascun array sono centrati; per ciascun gene calcoliamo il rapporto logaritmico medio di 20 pazienti. Per ciascun gene in ciascun paziente sottraiamo il rapporto logaritmico centrato dal rapporto logaritmico medio. La deviazione standard di questa distribuzione è 0.60. Questo corrisponde ad una variabilità del 44%. È piuttosto comune che, a causa delle differenze tra individui, ci siano consistenti sorgenti di variabilità. Questa è una della ragioni per cui è essenziale replicare gli esperimenti con parecchi individui. Il numero di individui necessari dipende dal tipo di esperimento che si sta sviluppando oltre che dal livello di variabilità della popolazione.

Metodi per stimare quanti replicati si debbano usare, sono discussi in profondità nel Capitolo 10.

Riassunto dei punti chiave

- Esprimenti con microarrays hanno parecchie sorgenti di variabilità.
- La variabilità può essere misurata e quantificata con esperimenti di calibrazione e studi pilota.
- Con l'assunzione log-normale, la variabilità può essere espressa come coefficiente percentuale.

Capitolo 7

Analisi dei geni espressi in Modo Differenziale

7.1 Introduzione

L'analisi dei dati è considerata la più grande e forse la più importante area della bioinformatica dei microarrays. Prendendo atto di ciò, vi sono tre capitoli in questo libro che descrivono i metodi per l'analisi dei dati, i quali nella loro totalità rispondono ai tre insiemi di questioni scientifiche inerenti i dati del microarray:

- Quali geni sono espressi in modo differenziale in un insieme di campioni, in relazione ad un altro insieme?
- Quali sono le relazioni tra i geni oppure tra i campioni che si stanno misurando?
- È possibile classificare i campioni sulla base delle misure di espressione dei geni?

In questo capitolo descriviamo i metodi posti dalla prima domanda: la ricerca di geni up-regolati e down-regolati;

I capitoli 8 e 9 rispondono alle altre due questioni. Questo capitolo copre una varietà di tecniche, e prende spunto sia dalla statistica classica che dalle teorie più moderne, per fornire un dettagliato resoconto di come si analizzino i dati dei microarrays a DNA per i geni espressi in modo differenziale. Iniziamo questo capitolo con tre esempi per illustrare cosa intendiamo per identificazione di geni espressi in modo differenziale.

Esempio 7.1: Insieme di dati 7A

I campioni sono presi da 20 pazienti con il cancro al seno, prima e dopo 16 settimane di trattamento medico con chemioterapia basata sul farmaco doxorubicin, ed analizzati usando i microarray. Desideriamo, dunque, identificare i geni che sono up-regolati e down-regolati nel cancro alla mammella, nell'ambito di questo trattamento medico[1].

Esempio 7.2: Insieme di dati 7B

Sono stati presi campioni del midollo osseo da 27 pazienti affetti da leucemia linfoblastica acuta (ALL) e da 11 pazienti affetti da leucemia mieloide acuta (AML) ed analizzati usando gli arrays Affimetrix. Desideriamo, in questo caso, identificare i geni che sono up-regolati e down-regolati in ALL relativamente ad AML [2].

[1] I dati sono presi dall'articolo di Perou ed altri (2000) e sono disponibili nello Stanford Microarray Database. Tutti I riferimenti sono forniti alla fine del capitolo.
[2] I dati sono presi dall'articolo di Golub ed altri (1999)) e sono disponibili nello Stanford Microarray Database

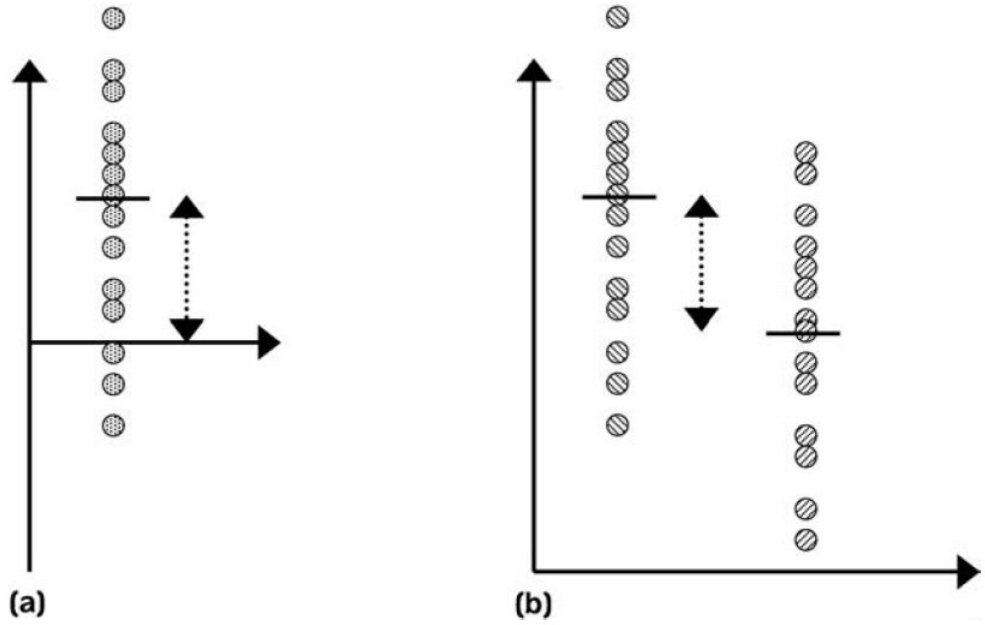


Figura 7.1: Dati appaiati e non appaiati. **(a)** Dati appaiati; le misure provenienti da ciascun individuo sono sottratte per produrre una singola misura. Vogliamo sapere se la media (o la mediana) delle misure sia, oppure no, differente da 0. **(b)** Dati non appaiati; c'è una misura di espressione del gene proveniente da ciascun individuo da ciascuno dei due gruppi. Vogliamo sapere se/oppure no, la media (o mediana) dei due gruppi è differente l'una dall'altra.

Esempio 7.3: Insieme di dati 7C

Vi sono quattro tipi di tumori nella prima infanzia caratterizzati da piccole cellule rotonde di colore blu: neuroblastoma (NB), linfoma non-Hodgkin (NHL), rhabdomyosarcoma (RMS) and tumori di Ewing (EWS). Sessantatre campioni di questi tumori, 12, 8, 20 e 23 in ciascuno dei gruppi, rispettivamente, sono stati ibridizzati ai microarray[3]. Vogliamo identificare i geni che sono espressi in modo differenziale in uno, o in più di uno, di questi 4 gruppi.

Ribadiamo il concetto: in tutti questi esempi siamo interessati ad identificare i geni espressi in modo differenziale.

I metodi che descriviamo in questo capitolo sono progettati per considerare un gene alla volta, al fine di determinare se/oppure no, esso è espresso in modo differenziale. Il metodo potrebbe, quindi, essere applicato ad ogni gene sul microarray con lo scopo di identificare quei geni che sono espressi in modo differenziale. Stiamo usando, dunque, il microarray come tool per studiare parecchi geni in parallelo -geni provenienti da molti individui.

Questo approccio contrasta con i metodi di analisi dei Capitoli 8 e 9, che sono specificamente indirizzati alle interazioni tra i geni sul microarray. Benché questi esempi siano simili, ciascuno di essi è un esempio di dati appaiati, dati disappaiati, oppure dati di struttura piú complessa. L'insieme di dati 7A è un esempio di dati appaiati (Figura 7.1a). Vi sono due misure per ciascun paziente: una prima del trattamento medico, e l'altra dopo il trattamento. Queste due misure sono correlate l'una all'altra; in sostanza, siamo

[3] I dati provengono dall'articolo di Khan e altri (2001) e sono disponibili presso lo Stanford Microarray Database

interessati alla differenza delle due misure (il rapporto logaritmico) per determinare se/oppure no, il gene è stato up-regolato o down-regolato nel corso del trattamento.

L'insieme dei dati 7B è un esempio di dati non appaiati (Figura 7.1b). Vi sono due gruppi di pazienti; noi siamo interessati a vedere se un gene è differenzialmente espresso tra i due gruppi. Non c'è una relazione intrinseca tra i pazienti di un gruppo ed i pazienti dell'altro gruppo. I dati appaiati e non appaiati richiedono una analisi leggermente differente; questo aspetto sarà delucidato nel corso del capitolo. L'insieme dei dati 7C ha quattro gruppi e richiede una analisi più complessa. In questo capitolo forniremo soltanto una breve introduzione ai tipi di dati più complessi, ed alle analisi delle varianze di essi (ANOVA): analisi, appunto, che sono di importanza essenziale, considerata la complessitá di questi dati. Il resto del capitolo è organizzato nei seguenti sei paragrafi:

Paragrafo 7.2: Concetti Fondamentali, introduce le idee che sono alla base di tutti i metodi descritti in questo capitolo: inferenza statistica, tests di ipotesi statistica, p-values ed indipendenza statistica.

Paragrafo 7.3: Statistica Parametrica Classica , *t*-Tests, discute sull'approccio statistico tradizionale all'analisi dei dati.

Paragrafo 7.4: Statistica Non Parametrica, rivolge l'attenzione a quei metodi che permettono una analisi robusta dei dati che non sono normalmente distribuiti.

Esamineremo sia la statistica non-parametrica tradizionale, sia il più moderno approccio del bootstrapping, che offre una combinazione tra la potenzialità del *t*-Tests e la robustezza della statistica non-parametrica tradizionale.

Paragrafo 7.5: Molteplicità del Testing, descrive i problemi statistici associati all'applicazione dei metodi di analisi a molti geni; inoltre, descrive un semplice metodo per risolvere questi problemi.

Paragrafo 7.6: ANOVA e General Linear Models (Modelli Generali di Analisi Lineare) fornisce una breve introduzione all'analisi dei dati piú complessi, come i dati dell'insieme 7C dove vi possono essere più di due gruppi di pazienti, oppure dove sono presenti insiemi di dati provenienti da ambienti in cui vi siano parecchi fattori che determinano le misure della espressione del gene.

7.2 Concetti fondamentali

Questo paragrafo illustra quattro concetti su cui si basano tutti i metodi descritti in questo capitolo:

- Inferenza Statistica
- Test di Ipotesi
- *p*-Values
- Indipendenza Statistica

Inferenza Statistica

L'Inferenza statistica costituisce il nucleo sia per le scienze che per la statistica classica. Consideriamo ancora l'insieme dei dati 7A. Siamo interessati ad identificare i geni che sono up-regolati oppure down-regolati nel cancro della mammella durante il corso della chemioterapia. Supponiamo che si sia scelto di sviluppare un esperimento su tutti i pazienti del mondo -affetti dal cancro alla mammella- che abbiano ricevuto questa cura. Tutto ciò descriverebbe in modo completo i risultati di questa terapia, ma sarebbe un esperimento lento e costoso: praticamente impossibile. Invece, abbiamo scelto un

campione4 di 20 pazienti, che confidiamo sia rappresentativo della popolazione con cancro alla mammella, e sviluppiamo l'analisi su questi pazienti. Quindi, ciò che facciamo, è tentare di generalizzare i risultati di questi 20 pazienti, facendo delle asserzioni scientifiche circa i cambiamenti dell'espressione del gene nel corso della chemioterapia del cancro alla mammella, estendendole all'intera popolazione dei pazienti con questa patologia (Figura 7.2). Stiamo tentando di fare una inferenza statistica: siamo esplicitamente interessati alle variabilità tra gli individui nella popolazione alla quale stiamo tentando di estrapolare l'analisi. Noi vogliamo catturare quanto più è possibile questa variabilità nel nostro esperimento e nelle analisi statistiche; pertanto, tentiamo di massimizzare l'utilizzo dei replicati biologici, dovendo ottemperare a problemi di budget ed a condizionamenti di ordine pratico. Possiamo anche tentare di includere sottotipi della popolazione, per esempio, l'età oppure fattori genetici, sia nel progetto dell'esperimento che nelle analisi statistiche.

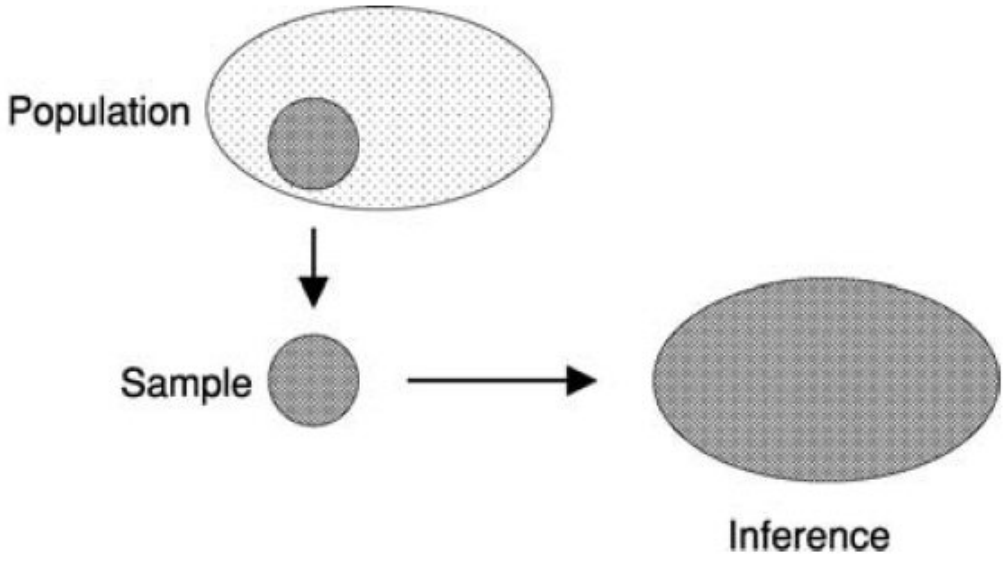


Figura 7.2: L'idea della inferenza statistica é centrale per la statistica classica e costituisce la base del metodo scientifico. Noi siamo interessati a descrivere una popolazione di individui; tuttavia, non è facile misurare ogni individuo della popolazione. Invece, scegliamo dalla popolazione un campione rappresentativo di individui e facciamo le misure su di essi. Quindi eseguiamo una estrapolazione delle misure degli individui per fare delle asserzioni circa la popolazione dalla quale il campione è derivato.

Test di Ipotesi e *p*-Values

Per ciascun gene nel'insieme dei dati 7A, abbiamo misure dell'espressione sia prima che dopo il trattamento medico di ciascun paziente. Queste misure dovrebbero essere normalizzate, memorizzate e trasformate in rapporto logaritmico per ciascun paziente, il quale descrive numericamente il campo di azione in cui il gene è differenzialmente espresso, e se esso è up-regolato o down-regolato. Vogliamo identificare quei geni che sono differenzialmente regolati in modo consistente nel'insieme dei 20 pazienti, con una attenzione particolare nell'asserire che questi geni siano differenzialmente espressi nel

[4] La parola campione qui si riferisce ad un campione statistico da una popolazione, e non ad un campione biologico che sarà ibridizzato ad un microarray. Sfortunatamente, la comunità degli studiosi di statistica e la comunità degli studiosi in biologia usano la stessa parola per riferirsi a differenti argomenti, che qualche volta conducono a confusione

corso della terapia. Nei primi esperimenti a microarray, i ricercatori avrebbero scelto una soglia, per esempio un'espressione differenziale di un fattore 2 , ed avrebbero selezionato quei geni la cui espressione differenziale media fosse stata maggiore della soglia. Da un punto di vista statistico, tuttavia, questo non è affatto un buon approccio per le seguenti due ragioni:

- Il rapporto di espressione medio non tiene conto del campo di azione in cui le misure dell'espressione differenziale del gene variano tra gli individui che si stanno studiando. Il livello della variabilità della popolazione sarebbe critico se noi tentassimo di usare il campione sperimentale per inferire una asserzione generale circa la popolazione *in toto*.
- Il rapporto di espressione medio non tiene conto del numero dei pazienti compresi nello studio, che gli statistici chiamano *ampiezza del campione*. Intuitivamente, ci aspetteremmo che tanto più grande è l'ampiezza del campione, quanto più alta la confidenza con cui potremmo determinare i geni che sono differenzialmente espressi.

Per queste ragioni, gli statistici determinano se/oppure no, un gene è espresso in modo differenziale, sulla base di metodologie conosciute con il nome di Test di Ipotesi. Un test di ipotesi costruisce un modello probabilistico per i dati osservati, basato su quella che viene comunemente chiamata ipotesi nulla. L'ipotesi nulla è l' ipotesi in cui non vi è alcun effetto biologico reale. Per un gene dell'insieme di dati 7A, ad esempio, l'ipotesi nulla sarebbe che questo gene non fosse differenzialmente espresso nel corso della chemioterapia con doxorubicin; per un gene nell'insieme di dati 7B, dovrebbe accadere che questo gene non sia differenzialmente espresso tra i pazienti ALL e AML. Se l'ipotesi nulla fosse vera, allora la variabilità dei dati non dovrebbe rappresentare l'effetto biologico che stiamo studiando, ma piuttosto i risultati dalle differenze fra gli individui, oppure errori di misura.

Ciascun test di ipotesi in questo capitolo costruisce un modello probabilistico soggiacente ad una ipotesi nulla. Usando questo modello, é possibile calcolare la probabilità di osservare una grandezza statistica - per esempio una espressione differenziale media di un gene- che rappresenta, al più, la possibilità di osservare un certo comportamento statistico nei dati. Questa probabilità è conosciuta come un *p*-value. Più piccolo è il valore del *p*-value, più bassa è la probabilità che i dati osservati abbiano una occorrenza casuale, e più significativo è il risultato. Per esempio, un *p*-value di 0.01 dovrebbe significare che vi è l'1% di possibilità di osservare almeno questo livello di espressione differenziale del gene come accadimento casuale.

Noi quindi selezioniamo i geni espressi in modo differenziale non sulla base dell'entità del rapporto di espressione tra condizione e controllo, ma sulla base del loro *p*-value. Inoltre ipotizziamo che sia molto improbabile che l'espressione differenziale osservata in quei geni con *p*-value molto bassi sia stata osservata per caso e che sia quindi più probabilmente dovuta ad un effetto biologico. Nella applicazioni tradizionali, un *p*-value inferiore a 0.01 dovrebbe essere stimato come risultato significativo. Nella applicazione dei microarrays, è necessario usare un *p*-value più stringente; questo è discusso nel dettaglio nel paragrafo 7.5.

Indipendenza Statistica

Due misure sono indipendenti statisticamente se la conoscenza del valore di una misura non fornisce alcuna informazione circa il valore dell'altra. Tutti i test statistici che descriviamo in questo capitolo, richiedono che le misure che stiamo analizzando siano statisticamente indipendenti. Tuttavia, misure replicate dallo stesso paziente, per esempio

spots replicati su un array, non sono statisticamente indipendenti e non potrebbero essere incluse come variabili separate nei test di ipotesi. Il metodo più semplice per essere sicuri che tutti i punti (facenti parte di un data set) dell'analisi siano indipendenti, è di combinare le misure non indipendenti dentro una singola variabile. Nel data set 7A, noi combiniamo due punti dallo stesso paziente sottraendo uno dall'altro in modo da creare un singolo punto (sempre appartenente a quel data set) per ciascun paziente e sviluppiamo l'analisi su questi valori. Noi potremmo non trattare i 40 campioni come variabili indipendenti.

7.3 Statistica Classica Parametrica - *t*-TESTS

Questi test di ipotesi costituiscono storicamente l'approccio standard per analizzare i dati nelle forme dei data set 7A o 7B. Vi sono due versioni del test: ciò dipende dal fatto se i dati siano appaiati oppure no.

t-Test per Dati Appaiati o a un solo Campione

Il *t*-Test per dati appaiati, conosciuto anche dai statistici come *t*-Test a un solo campione[5], è applicabile ai dati che sono nella forma del data set 7A. In questo data set, c'è una coppia di misure per ciascun paziente, una prima del trattamento medico, e l'altra dopo il trattamento medico, e queste sono combinate per generare un solo rapporto logaritmico per ciascun paziente. Quindi, i dati da analizzare appaiono come una singola colonna di numeri, una per ciascun paziente. Da questo set di dati, si potrebbe calcolare la statistica-*t*, usando la seguente formula:

$$t = \frac{\bar{x}}{s/\sqrt{n}}$$

Equazione 7.1

Dove $\bar{x}$ è la media dei rapporti logaritmici di ciascuno dei pazienti; ***s*** è la deviazione standard del campione dei pazienti; ***n*** è il numero dei pazienti nel'esperimento. Viene quindi calcolato un *p*-value dalla statistica-*t* confrontando questa con la distribuzione-*t* con un numero appropriato di **gradi di libertà.** Il grado di libertà è il numero di variabili indipendenti dell'analisi. Nel caso di t-test appaiati, esso è uguale al numero di pazienti meno 1.

Il metodo di comparazione del rapporto logaritmico medio con una soglia per determinare i geni differenzialmente espressi, dovrebbe fare uso esattamente della quantità $\bar{x}$,dell'equazione 7.1. Tuttavia, dobbiamo osservare che il *t*-Test è molto più sofisticato rispetto a questo metodo. Il significato di geni espressi in modo differenziale dipende non soltanto dal rapporto logaritmico medio, ma dipende anche -sia dalla variabilità della popolazione, sia dal numero di individui inclusi nello studio. Così un gene potrebbe essere differenzialmente espresso solo 1,5 volte, ma indicato come significativo se la variabilità della popolazione fosse piccola. Similmente, più sono numerosi gli individui in un esperimento, più facile è la determinazione di geni espressi in modo differenziale.

[5] La confusione derivante dal'uso duplice della parola campione, è perfino maggiore quando si riferisce ai nomi del t-test: un t-test ad un campione è usato quando due campioni biologici sono presi da ciascun paziente (dove vi è un campione statistico di pazienti); *t*-Test a due campioni sono usati quando un campione biologico è preso da ciascun paziente (quando vi sono due campioni statistici del paziente). Io penso che sia veramente chiaro il beneficio di avere biologi e statistici che lavorano insieme in stretto rapporto perché l'uno impari il linguaggio dell'altro, e viceversa.

I *t*-Test appaiati sono ampiamente implementati nel software dei computer, incluso Excel, SPSS, SAS, S+, R e GeneSpring. Questi test sono anche implementati direttamente nel codice (ndt.: le funzioni di calcolo statistico sono in codice oggetto, e vengono richiamate dalle librerie al momento della compilazione).

Esistono, infatti, librerie disponibili per i linguaggi di programmazione più diffusi. La maggior parte degli utilizzatori usa questo software: per esempio, in Excel si utilizza la funzione TTEST, ed in R si utilizza la funzione *t*-Test.

TABLE 7.1: Data for ACAT2 from Data Set 7A

Patient	Before Treatment	After Treatment	Log Ratio	Fold Difference
7	−0.86	−2.17	−1.30	−2.47
10	−1.97	−1.93	0.04	+1.03
12	−2.07	−1.28	0.79	+1.73
14	−1.91	−2.32	−0.41	−1.33
15	−0.94	−2.00	−1.06	−2.09
18	−1.29	−1.74	−0.45	−1.37
26	−1.09	−1.54	−0.44	−1.36
27	−0.65	−0.60	0.06	+1.04
39	−1.69	−2.06	−0.37	−1.30
41	−0.79	−1.22	−0.43	−1.35
47	−1.19	−2.11	−0.91	−1.88
48	−1.36	−1.40	−0.04	−1.03
53	−1.11	−1.59	−0.48	−1.40
61	−1.82	−1.72	0.10	+1.07
100	−2.22	−2.13	0.10	+1.07
101	−1.76	−1.94	−0.18	−1.14
102	−1.51	−2.37	−0.86	−1.81
104	−1.65	−1.98	−0.33	−1.25
109	−0.78	−1.49	−0.71	−1.63
112	−1.80	−1.82	−0.03	−1.02
Average	**−1.42**	**−1.77**	**−0.35**	**−1.21**
Sample SD	**0.48**	**0.43**	**0.48**	

Note: In questo esperimento, i campioni provenienti da prima e dopo il trattamento sono stati ibridizzati a due arrays separati, con un campione comune di riferimento sul secondo canale. Le misure, prima e dopo il trattamento, sono il logaritmo dei rapporti dei campioni sperimentali rispetto ai campioni di riferimento. Il logaritmo del rapporto è la differenza tra questi due valori. I logaritmi sono presi in base 2, così che il valore di 1 rappresenta una regolazione 2-fold up, ed il valore -1 rappresenta una regolazione 2-fold down. Le deviazioni standard del campione sono state calcolate con un denominatore di n-1 =19 per assicurare che gli stimatori della popolazione non siano polarizzati.

Esempio 7.4: *t*-Test appaiato applicato al gene dal data set 7A

Il gene acetyl-Coenzima A acetyltransferasi 2 (ACAT2) è presente sui microarray usati per i dati del cancro alla mammella, data set 7A. Noi possiamo usare un t-test appaiato per determinare se/oppure no, il gene è espresso in modo differenziale nel corso del trattamento medico della chemioterapia basata sul farmaco doxorubicin (Tabella 7.1). In questo particolare esperimento, i campioni prima e dopo la chemioterapia sono stati

ibridizzati a microarray separati, con un campione di riferimento sull'altro canale. Vi sono tre passi:

1. Normalizzare i dati usando uno dei metodi descritti al capitolo 5; questo non è parte del *t*-Test ma è di solito richiesto prima che i dati del microarray possano essere analizzati. Siccome si tratta di un esperimento con campione di riferimento, noi calcoliamo il rapporto logaritmico del campione sperimentale, relativo al campione di riferimento,sia prima che dopo il trattamento medico in ciascun paziente.
2. Calcolare un singolo rapporto logaritmico per ciascun paziente che rappresenta la differenza nell'espressione del gene dovuta al trattamento medico, sottraendo il rapporto logaritmico per il gene prima del trattamento, dal rapporto logaritmico del gene dopo il trattamento.
3. Sviluppare il t-test con l'ausilio di un pacchetto software, oppure calcolando la media e la deviazione standard del campione dei rapporti logaritmici ed applicando l'equazione 7.1. Il t-statistic è 3.22; questo è comparato con un t-distribuzione di 19 gradi di libertà (20 -1). Il *p*-value per un campione a due code t-test è 0.0045, che è significativo ad un livello di confidenza dell'1%.

Noi potremmo concludere, per il momento, che il gene è stato down-regolato ad un livello di confidenza dell'1%. Vi sono parecchie opposizioni a questa affermazione che diventeranno evidenti nel'ultima parte di questo capitolo.

t-Test per Dati non appaiati, oppure *t*-Test per due Campioni

Il *t*-Test non appaiato è molto simile al t-test appaiato, essendo la differenza relativa al progetto sperimentale. Nel data set 7A, i dati sono appaiati: vi sono due valori per ciascun paziente, che sono prima sottratti l'uno dall'altro, e quindi il *t*-Test appaiato è applicato alla differenza. Nel data set 7B, i dati sono non appaiati: vi sono due gruppi di pazienti in cui non vi è alcuna relazione con ciascun altro, ed effettuiamo il test per vedere se la differenza della media tra i due gruppi è zero. Il *t*-Test non appaiato, denominato anche *t*-Test a due campioni, usa una formula simile al *t*-Test appaiato:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}$$

Equazione 7.2

dove $\bar{x}_1$ ed $\bar{x}_2$, sono le medie dei due gruppi; s_1 ed s_2 sono le deviazioni standard del campione dei due gruppi; e n_1 ed n_2 sono le grandezze dei due gruppi. Vi sono al momento due forme per il *t*-Test non appaiato: la versione quì adottata permette alla deviazione standard tra i due gruppi di essere differente, e si tratta di una migliore versione per l'uso con i microarray. Vi è anche una versione che calcola una singola deviazione standard per tutti i dati. La statistica-*t* è confrontata con una t-distribuzione con un numero appropriato di gradi di libertà, per ottenere il *p*-value[6].

Il *t*-Test non appaiato è implementato nello stesso range del *t*-Test appaiato. Grazie a questa similarità tra i due test, il software usa frequentemente la stesa formula, per esempio la formula TTEST in Excel oppure la formula *t*-Test in R.

[6] Il numero di gradi di libertà per due campioni t-test con una varianza disuguale è dato da una formula molto complicata; il lettore interessato è rimandato ad uno dei due libri di statistica riportati alla fine del capitolo.

TABLE 7.2: Data for Metallothionein IB from Data Set 7B

Patient	ALL Log	Patient	AML Log
1	8.60	28	8.42
2	7.85	29	8.35
3	8.85	30	9.58
4	8.20	31	9.18
5	7.60	32	9.41
6	8.21	33	8.96
7	8.47	34	8.81
8	8.51	35	9.55
9	8.75	36	8.18
10	6.75	37	8.71
11	7.93	38	9.46
12	7.71		
13	7.88		
14	7.55		
15	6.61		
16	8.75		
17	9.32		
18	8.40		
19	7.16		
20	8.41		
21	4.75		
22	7.92		
23	7.82		
24	8.42		
25	7.08		
26	7.38		
27	9.29		
Average	7.93		8.97
Sample s.d.	0.94		0.51
Fold Ratio	−1.84		+1.84

Nota: Questi dati provengono dagli arrays Affimetrix; i valori sono stati registrati (in base 2) per essere sicuri che i dati fossero normalmente distribuiti.

Esempio 7.5: *t*-Test non appaiati applicato al gene dal data set 7B

Il gene della metallotionina IB è presente sull'array Affimetrix usato per i dati di leucemia, il data set 7B. Noi vogliamo identificare se/oppure no, questo gene è espresso in modo differenziale tra i pazienti AML ed ALL. Vogliamo, cioè, identificare i geni che sono up-regolati oppure down-regolati in AML in relazione ad ALL (Tabella 7.2). Vi sono tre passi da seguire:

1. I dati sono log trasformati (vedi Capitolo 5).
2. La media e le deviazione standard del campione sono calcolate per ciascun insieme di pazienti: una media ed una deviazione standard per tutti i pazienti ALL, ed una media separata ed una deviazione standard per i pazienti AML. Il test statistico determina se queste due medie sono uguali.

3. Il t-test è calcolato usando uno dei pacchetti software, oppure per mezzo della equazione 7.2. In questo caso, il t-statistic è 4.35. Questo è confrontato con una t-distribuzione con 33 gradi di libertà e produce un *p*-value di 0.00012.

Possiamo concludere che l'espressione della metallotionina IB è significativamente più alta in AML piuttosto che in ALL a livello dell'1%.

Requisiti dei *t*-Tests

I *t*-Tests appena descritti sono usati in statistica molto frequentemente, e altrettanto frequentemente appaiono nella letteratura medica e biologica. Tuttavia, i *t*-Test richiedono che la distribuzione dei dati che si sta esaminando sia normale. Questo ha qualche piccola differenza di significato per i tests appaiati e non appaiati:

- Per *t*-Tests appaiati, è la distribuzione dei dati sottratti che deve essere normale.
- Per i *t*-Tests non appaiati, la distribuzione di entrambi i data set deve essere normale.

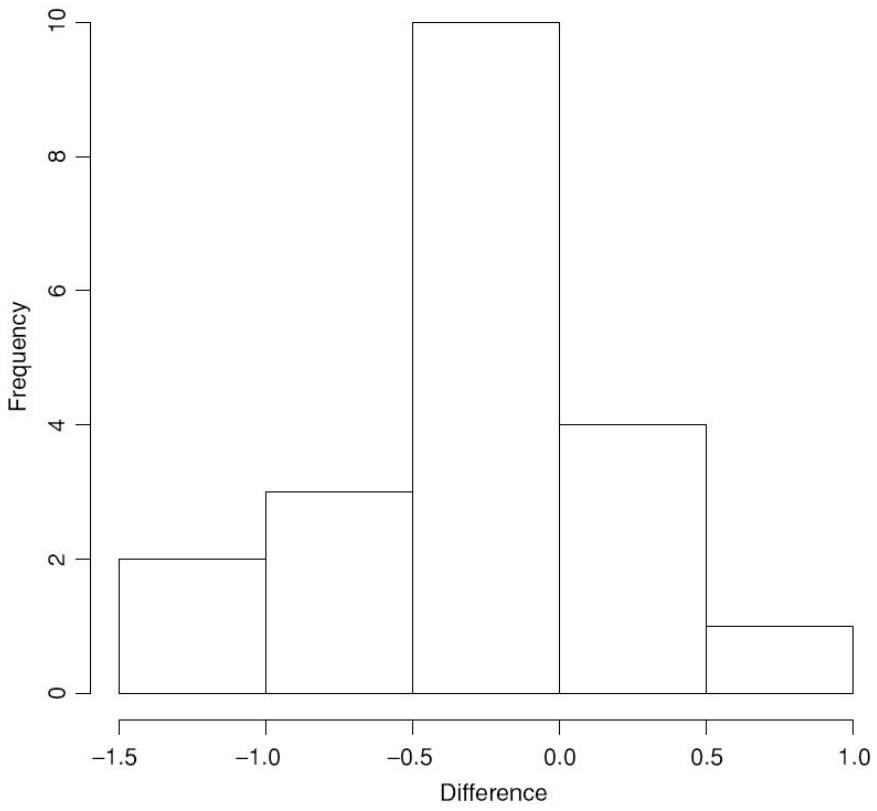


Figura 7.3: Istogramma della differenza tra i rapporti logaritmici dell'espressione di ACAT2 in 20 pazienti con il cancro alla mammella prima e dopo un trattamento medico di chemioterapia basata sul farmaco doxorubicin. I dati sono approssimativamente normali; la media delle distribuzioni appare essere inferiore allo zero, suggerendo che questo gene potrebbe essere down-regolato.

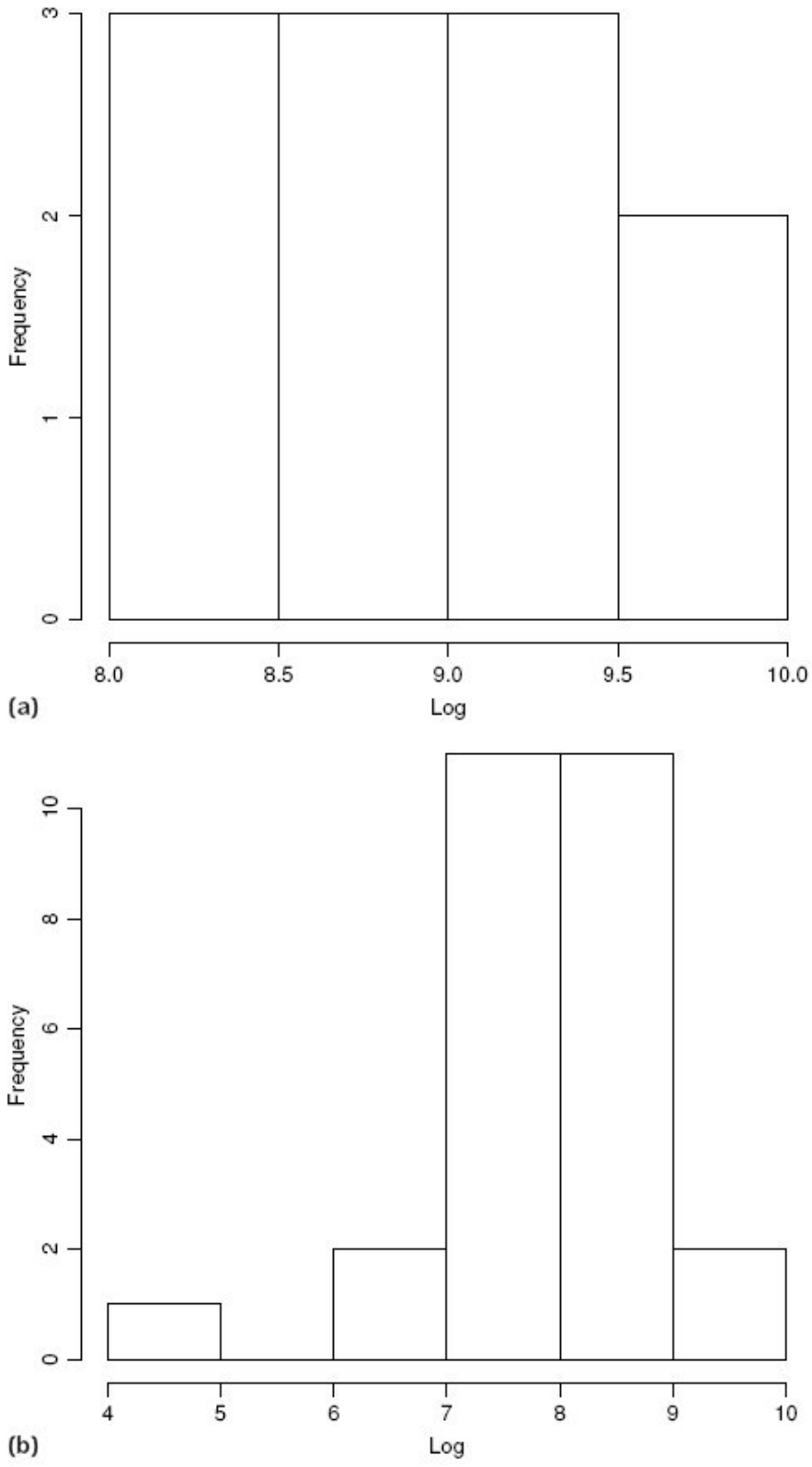


Figura 7.4: Istogramma del logaritmo dell'espressione del gene della metallotionina in (a) 11 pazienti AML e (b) 27 pazienti ALL. Entrambe le distribuzioni sono approssimativamente normali. La media dell'istogramma per tutti i pazienti ALL appare essere piú bassa che la media dei pazienti AML, suggerendo che questo gene potrebbe essere regolato in modo differenziale in queste due patologie.

TABLE 7.3: Data for Diubiquitin from Data Set 7A

Patient	Unlogged Difference	Log Ratio	Fold Change
7	−10.08	−2.91	−7.54
10	0.85	0.62	+1.54
12	−0.28	−0.11	−1.08
14	0.04	0.08	+1.06
15	−0.68	−0.42	−1.34
18	0.17	0.12	+1.09
26	−4.93	−0.99	−1.99
27	−0.12	−0.16	−1.12
39	−1.67	−0.44	−1.35
41	−27.98	−1.64	−3.12
47	−0.92	−0.55	−1.46
48	−2.00	−0.99	−1.99
53	−3.04	−1.37	−2.58
61	−3.80	−2.05	−4.14
100	−3.53	−3.20	−9.18
101	−1.44	−1.12	−2.17
102	−0.62	−0.72	−1.64
104	−4.50	−1.19	−2.27
109	−0.23	−0.34	−1.27
112	0.10	0.12	+1.09

Nota: La differenza non logaritmica include due valori estremi, il paziente 7 ed il paziente 41. Questo ha un effetto detrimentale sull'analisi dei dati, e produce risultati inaffidabili nel *t*-Test applicato ai dati sorgenti. Questi punti non sono valori estremi nei dati logaritmici, e quindi il *t*-Test applicato ai dati logaritmici è più affidabile.

Esempio 7.6: Esempi di dati normalmente distribuiti

La figura 7.3 mostra un istogramma dei dati dall'Esempio 7.4, relativo al gene ACAT2 nel set di dati 7A. Benché il campione sia piccolo, i dati sembrano approssimativamente normali. La figura 7.4 mostra l'istogramma dei dati dall'esempio 7.5, il gene della metallotionina nel set di dati 7B. Ancora, i campioni sono piccoli, particolarmente il campione AML che ha soltanto 11 pazienti, ma i dati appaiono approssimativamente normali. Entrambi questi due data sets soddisfano il requisito di normalità, così i *t*-Tests sono analisi appropriate su questi dati.

Esempio 7.7: Esempi di dati che non sono normalmente distribuiti

I dati sorgenti per il gene per la diubiquitina non sono normalmente distribuiti (Figura 7.5a; Tabella 7.3). Vi sono due dati estremi, con valore approssimativamente di -10 e -28. Quando il *t*-Test è applicato ai dati non logaritmici, il *p*-value è di 0.03, che non è significativo al livello dell'1%.

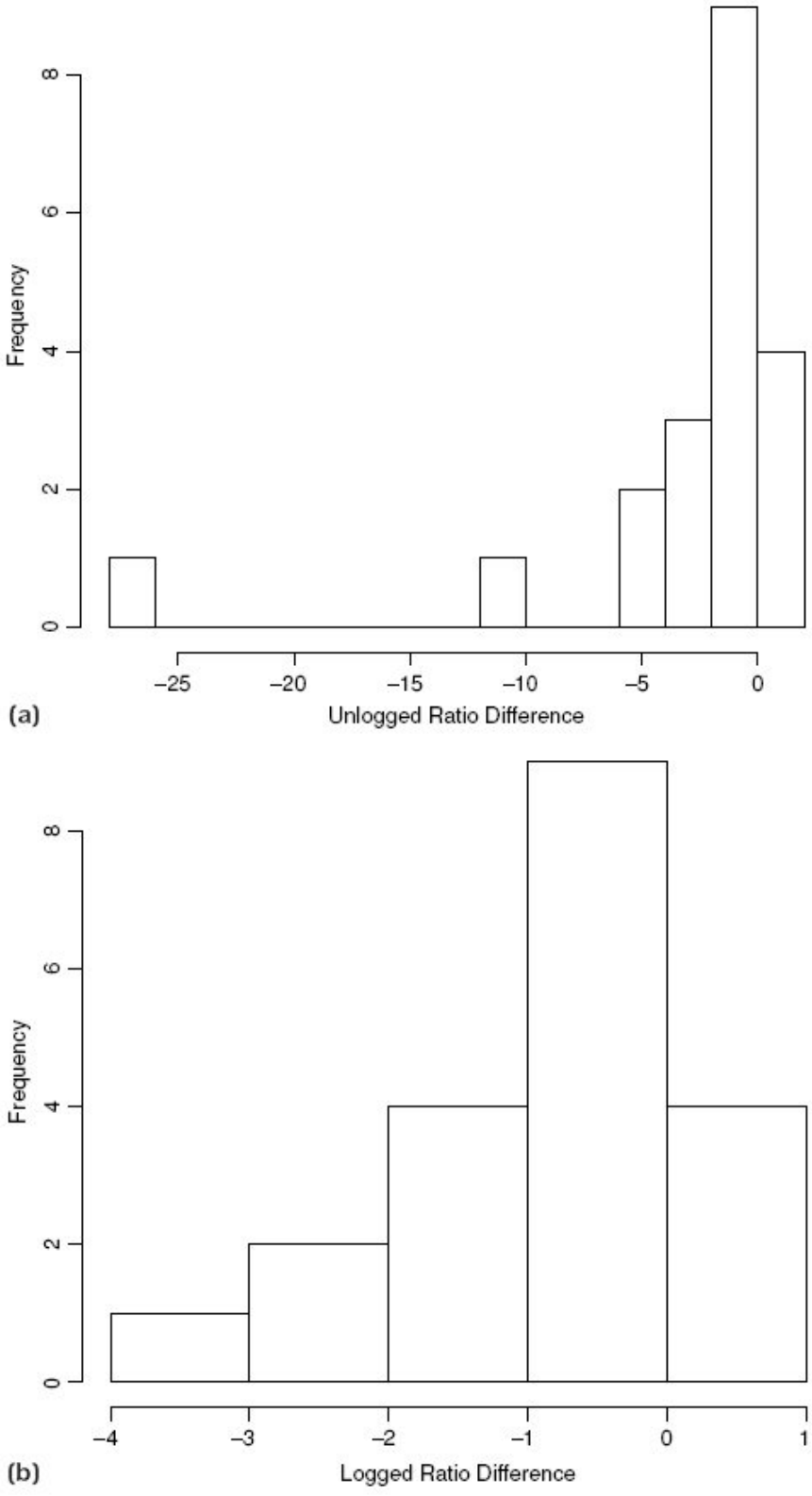


Figura 7.5: Istogrammi delle differenze di espressione del gene della diubiquitina in 20 pazienti con cancro alla mammella. **(a)** I dati non sono stati trasformati in logaritmi. La distribuzione non è normale: vi sono due elementi fuori posto, con valori approssimativamente di -20 e -11. **(b)** I dati sono stati memorizzati. La distribuzione è normale; gli elementi fuori posto sono stati eliminati. Si noti che in entrambi i casi la differenza dalla media è inferiore allo zero. Tuttavia, con i dati non logaritmici, un *t*-Test fornisce un risultato non significativo poiché l'errore standard della media è così alto, mentre per i dati logaritmici , un *t*-Test è significativo poiché l'errore standard è molto minore.

TABLE 7.4: Summary Statistics for Diubiquitin Gene from Data Set 7A

Data	Mean	Standard Deviation	Standard Error	Standard Error/Mean	*t*-statistic	*p*-value
Unlogged	−3.23	6.36	1.46	−0.45	−2.27	0.03
Logged	−0.86	1.00	0.23	−0.27	−3.86	0.001

Nota: L'errore standard dei dati non logaritmici è relativamente grande in relazione alla media; ciò a causa dei valori dei due elementi fuori posto. A causa di ciò, il *p*-value non è significativo. I dati registrati sono più vicini ad una distribuzione normale; gli elementi fuori posto non hanno valori estremi, e l'errore standard è piccolo in relazione alla media. Ciò si riflette nel *p*-value, che è significativo.

Dopo che i dati sono stati trasformati in logaritmi, i dati risultanti non hanno piú valori estremi, e la distribuzione appare normale (Figura 7.5b). Quando il *t*-Test è applicato ai data logaritmici, il *p*-value è di 0.001, che è significativo al livello dell'1%. Quando usiamo l'analisi corretta, concludiamo che il gene è significativamente down-regolato. Questo è in linea con quanto dovremmo aspettarci dall'ispezione della figura 7.5b; in molti pazienti, questo gene è down-regolato. Non avremmo raggiunta tale conclusione con l'analisi non corretta. Questo esempio è abbastanza contro intuitivo. Si poteva pensare che i dati non logaritmici, con i loro due elevati valori negativi, avrebbero mostrato una più chiara evidenza di down-regolazione rispetto ai dati logaritmici. In effetti, è vero l'inverso. Si può vedere il perché dal riassunto statistico (Tabella 7.4): i due valori estremi incrementano l'errore standard dei dati sorgenti. Questo errore standard così alto decrementa la certezza del test, che produce un risultato meno significativo. Al contrario, l'errore standard dei dati logaritmici é molto più piccolo relativamente alla media. Come risultato, la variabilità dei dati è più piccola, ed il *t*-Test produce un risultato significativo.

7.4 Statistica Non-Parametrica

Questo paragrafo discute i metodi che non si basano sull'assunzione che i dati siano normalmente distribuiti. Vi sono due buone ragioni per usare questi metodi anziché usare il *t*-Test per l'analisi dei dati del microarray.

- I Dati del microarray sono disturbati dal rumore. Vi sono molte sorgenti di variabilitá in un esperimento a microarray, e i valori estremi sono frequenti. Quindi la distribuzione delle intensità per molti geni non è una distribuzione normale. I metodi non parametrici sono robusti per far fronte alla presenza di valori estremi ed a dati disturbati dal rumore.
- L'Analisi dei dati dei microarray è **ad alto flusso.** Quando sviluppiamo un *t*-Test su un singolo dada set, è immediato verificare la distribuzione dei dati stessi per vedere se ottemperano ad una distribuzione normale.

Tuttavia, quando si analizzano molte migliaia di geni in un microarray, è necessario verificare la normalità di ogni gene per poter assicurare che il *t*-Test sia appropriato. Quei geni che avessero valori estremi oppure che fossero distribuiti non normalmente, dovrebbero richiedere una analisi differente. È molto più sensato applicare un test che non sia condizionato dalla distribuzione e quindi possa essere applicato alla totalità dei geni in un solo passo.

Vi sono due tipi di test non parametrici che discuteremo in questo paragrafo:

- Tests non parametrici classici equivalenti ai test parametrici ma che non sottostanno all'assunzione che i dati siano normalmente distribuiti.
- Test Bootstrap sono molto moderni ed applicabili ad un largo insieme di analisi.

Statistica Classica Non-Parametrica

Questi sono metodi semplici da applicare e sono implementati in tutti i pacchetti statistici, inclusi SPSS, SAS, S+ ed R, ma non in Excel. Vi sono equivalenti non parametrici sia per i test appaiati che per i test non appaiati descritti nel paragrafo 7.3. L'equivalente non-parametrico del test appaiato è denominato **Wilcoxon sign-rank test**. L'equivalente non-parametrico del *t*-Test non appaiato è denominato **Mann-Whitney test**, oppure -talvolta- **Wilcoxon rank-sum test**.

Il Wilcoxon sign-rank test lavora sostituendo il valore vero del rapporto logaritmico dei dati con la posizione in graduatoria (rank) in accordo al valore del rapporto logaritmico: 1 per i più piccoli, 2 per i secondi più piccoli, e così via. La somma dei valori di rank per i "valori" positivi (up-regolati) viene calcolata e confrontata al posto della tabella precompilata per ottenere il *p*-value. Il test di Mann-Whitney è simile. I dati provenienti dai due gruppi sono combinati e messi in graduatoria: 1 per i più piccoli, 2 per i secondi più piccoli, e così via. I valori di rank per il gruppo maggiore sono sommati e questo numero è confrontato con una tabella precalcolata per ottenere un *p*-value. Questi test hanno il vantaggio di non richiedere che i dati siano normalmente distribuiti, benché il sign-rank-test richieda che i dati siano simmetrici. Lo svantaggio di questi tests è che essi sono meno potenti dei loro equivalenti parametrici, o dei loro equivalenti boostrap. La potenza di un test statistico è definita come la probabilità di vedere un risultato positivo quando effettivamente esista un risultato positivo da vedere.

Discuteremo la potenza in maggiore dettaglio nel capitolo 10. A causa della perdita di potenza, la statistica classica non parametrica non è diventata popolare per l'uso con i dati dei microarrays; invece, di converso, i metodi bootstrap tendono ad essere preferiti.

Esempio 7.8 : Wilcoxon sign-rank test Diubiquitin dal data set 7A

Il Wilcoxon sign-rank test è applicato ad entrambe le versioni dei dati, registrati e non, di diubiquitina dall'esempio 7.7. I risultati sono come segue:

Non registrati: *p*-value è 0.00032
Registrati: *p*-value è 0.00048

In entrambi i casi, il test fornisce un risultato significativo, in accordo con l'analisi *t*-Test sui dati registrati. Il test Wilcoxon è robusto in presenza di valori estremi e fornisce un risultato significativo perfino sui dati non logaritmici.

Esempio 7.9: MANN-WHITNEY Test su recettore - like TYROSINA KINASE dal data set 7B

Il recettore del gene-tipo tirosina chinasi (RYK) appare sull'array Affimetrix usato per il data set 7B. Questi sono stati creati usando una versione primitiva della suite di analisi dei dati dei microarray della Affimetrix. Un certo numero di valori hanno un risultato negativo e sono stati sostituiti con zero.(vedi paragrafo 5.2) nei dati logaritmici del data set (tabella 7.5).

TABLE 7.5: Data for the Gene RYK from the Leukemia Data Set

Patient	ALL Log	Patient	AML Log
1	7.22	28	6.78
2	5.25	29	4.95
3	6.58	30	6.52
4	6.19	31	4.81
5	3.00	32	6.19
6	5.61	33	6.38
7	0.00	34	6.67
8	0.00	35	7.34
9	7.21	36	3.81
10	0.00	37	6.09
11	0.00	38	6.02
12	0.00		
13	6.81		
14	6.29		
15	1.00		
16	1.58		
17	7.29		
18	4.25		
19	5.32		
20	5.91		
21	4.58		
22	6.02		
23	6.00		
24	2.81		
25	5.78		
26	4.46		
27	0.00		
Average	4.60	Average	5.96

Nota: Questi dati furono generati dagli arrays Affimetrix usando la versione 4 del loro software. I geni per cui l'intensità di ibridizzazione delle sonde non adattate, furono più grandi dei risultati delle sonde verrebbero risultati negativi; questi furono sostituiti con lo zero nei dati registrati (vedere il paragrafo 5.2). L'ultima versione del software Affimetrix non genera piú numeri negativi.

Il *p*-value del Mann-Whitney test è 0.039, che non è significativo ad un livello di confidenza dell'1%. I due *p*-test del campione applicati a questi dati danno un *p*-value di 0.0032, che è significativo ad un livello di confidenza dell'1%.

Le risposte sono differenti poiché né i dati ALL né i dati AML sono normalmente distribuiti (Figura 7.6). Ancora peggio, l'insieme dei dati ALL (Figura 7.6A) è bimodale, con 10 pazienti che hanno una espressione molto debole, o addirittura non ne hanno affatto, e, di converso, i rimanenti pazienti mostrano relativamente una espressione molto alta. Il campione ALL è molto piccolo, e quindi è difficile trarre delle conclusioni, ma anche esso appare non essere distribuito normalmente.

Pertanto, il *t*-Test non è una analisi appropriata, e non dovremmo credere alla significatività dei risultati dal t-test. Potremmo concludere - in base all'analisi non parametrica - che questo gene non è espresso differenzialmente in modo significativo tra queste due patologie. Tuttavia, dobbiamo ricordare che il test Mann-Whitney è meno potente ed è molto probabile che conduca ad un falso negativo.

Nel prossimo paragrafo mostreremo il test bootstrap, applicato agli stessi dati e che fornisce, invece, un risultato significativo.

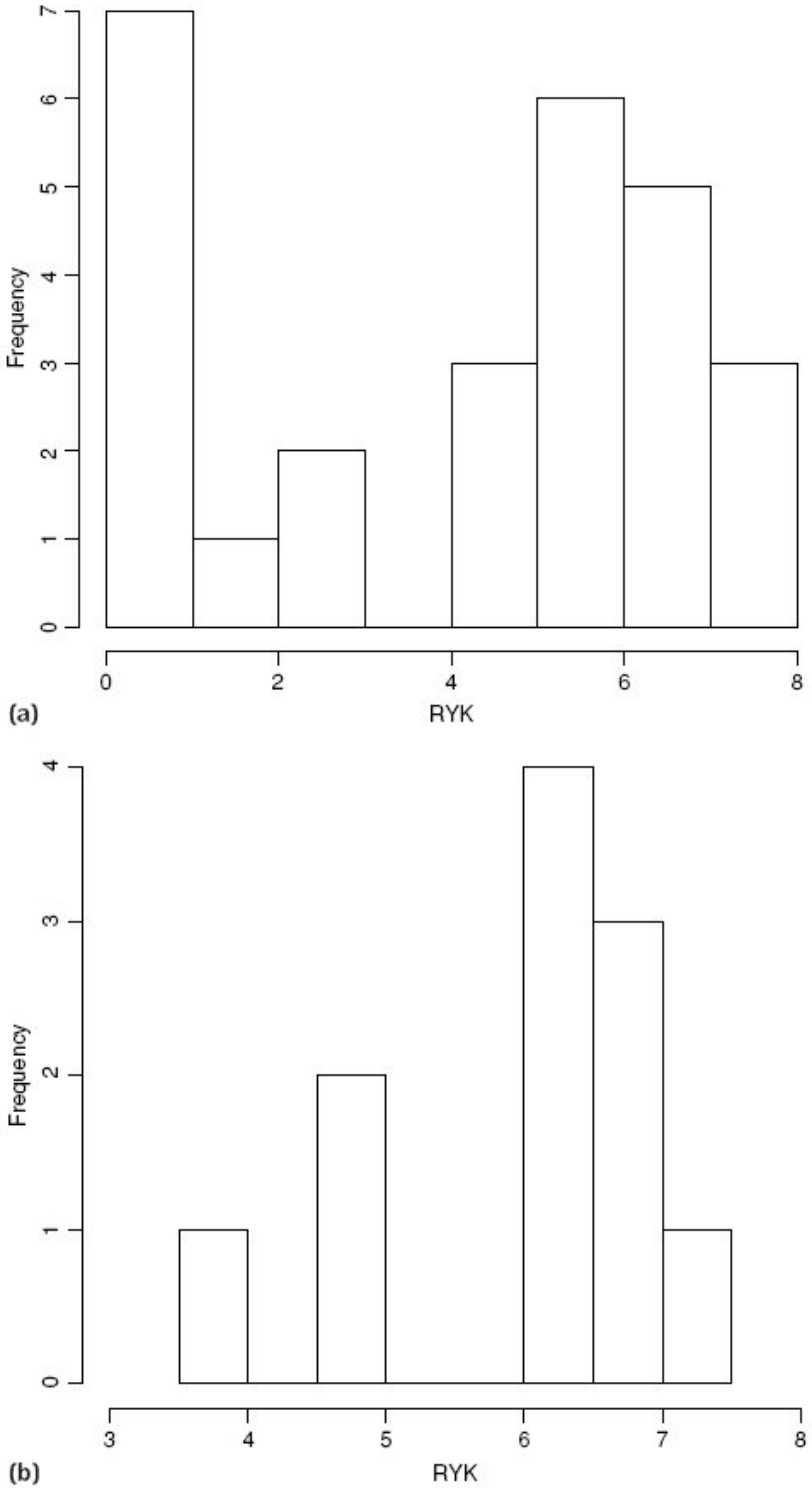


Figura 7.6: Istogramma del logaritmo della espressione del gene di RYK in (a) 27 pazienti ALL e **(b)** 11 pazienti AML. I risultati negativi nei dati sorgenti sono stati sostituiti con lo zero. Nessuna distribuzione è normale. I dati ALL sono bimodali, che significa presentano due picchi: 10 pazienti hanno un espressione debole, oppure non hanno espressione, e 17 pazienti presentano una buona espressione del gene. L'AML è un piccolo data set, ma è anch'esso non normale.

Analisi Bootstrap

Come con i tests classici non parametrici, l'analisi bootstrap non richiede che i dati siano normalmente distribuiti e sono, quindi, test robusti in presenza del rumore e degli artefatti sperimentali. Essi sono anche più potenti rispetto ai test non parametrici classici. Pertanto, le analisi bootstrap sono più adatte ad analizzare i microarrays, rispetto agli altri test, o ai test non parametrici classici (Tab. 7.6). Lo svantaggio dei metodi bootstrap è che sono moto pesanti da un punto di vista computazionale (computing intensive), così che la loro adozione è avvenuta solo con l'avvento dei moderni computers.

Esistono due metodi bootstrap equivalenti sia per l'analisi dei dati appaiati che per l'analisi dei dati non appaiati, come descritto precedentemente. È possibile usare i metodi bootstrap per analisi più complesse, come, ad esempio, i modelli ANOVA) e l'analisi dei clusters (Capitolo 8). In questo paragrafo, descriveremo come il bootstrap lavori sui dati non appaiati, poiché questo è il modo più semplice per comprenderlo. Alla fine del capitolo indichiamo come riferimento un libro eccellente sui metodi bootstrapping ove si desiderasse studiare ulteriormente questi metodi.

TABLE 7.6: Advantages and Disadvantages of Different Statistical Analyses

t-Tests	Non-parametric Tests	Bootstrap Analyses
✓ Easy	✓ Easy	✓ Robust
✓ Powerful	✓ Robust	✓ Powerful
✓ Widely implemented	✓ Widely implemented	× Requires use of specialist packages or programming
× Not appropriate for data with outliers	× Less powerful	

Con l'analisi dei dati non appaiati, vi sono due gruppi, e noi tentiamo di identificare se/oppure no, la media dei due gruppi sia differente. Per esempio, con il gene RYK dai dati del data set 7B, vi sono 27 misure dai pazienti ALL ed 11 misure dai pazienti AML; vogliamo conoscere se/oppure no, il gene sia espresso differenzialmente tra i due gruppi di pazienti. Sotto l'ipotesi nulla, non vi è alcuna differenza nel'espressione del gene tra i due gruppi. In questo caso qualsiasi data set misurato potrebbe essere osservato in ciascuno degli individui; nell'esempio, qualsiasi paziente AML potrebbe aver avuto qualsiasi delle 38 misure indicate nella tabella 7.5 associate ad entrambi i pazienti AML ed ai pazienti ALL. Il metodo bootstrap lavora costruendo un grande numero di data set aleatori ricampionando i dati originali, in cui a ciascun individuo viene assegnato aleatoriamente uno dei dati, che potrebbe provenire da uno dei due gruppi (Figura 7.7)[7].

Quindi il data set bootstrap appare come un set di dati reali, in quanto essi hanno valori simili, ma sono un non-senso biologico in quanto i valori sono stati resi aleatori. Lo spirito del test è quello di confrontare alcune proprietà dei dati reali con un distribuzione avente le stesse proprietà nel set aleatorio. La proprietà piú comunemente usata è il t-statistic (Equazione 7.2); questa è una buona misura poiché essa correla le differenze di

[7] Ci sono infatti due modi per sviluppare un metodo bootstrap: con o senza sostituzione. Quando il bootstrap è sviluppato con sostituzione, differenti individui nel dati del bootstrap potrebbero avere lo stesso valore dai dati reali. Quando il bootstrap è sviluppato senza sostituzione, ciascuno dei valori reali è utilizzato solo una volta nei dati del bootstrap. In questo capitolo, noi descriviamo il metodo con sostituzione, mentre il software di analisi di significatività (SAM) che menzioneremo più tardi, sviluppa un bootstrap senza sostituzione. Benché vi sia qualche dibattito su quale metodo sia il migliore, in parole povere i due metodi sono alquanto equivalenti e producono risultati molto simili.

espressione medie (fold ratio) alla variabilitá della popolazione ed al numero di individui nell'esperimento. Tuttavia, non usiamo la t-distribution per calcolare il *p*-value. Invece, noi generiamo una distribuzione empirica usando il t-statistic calcolato dai dati bootstrap che sono stati resi aleatori.

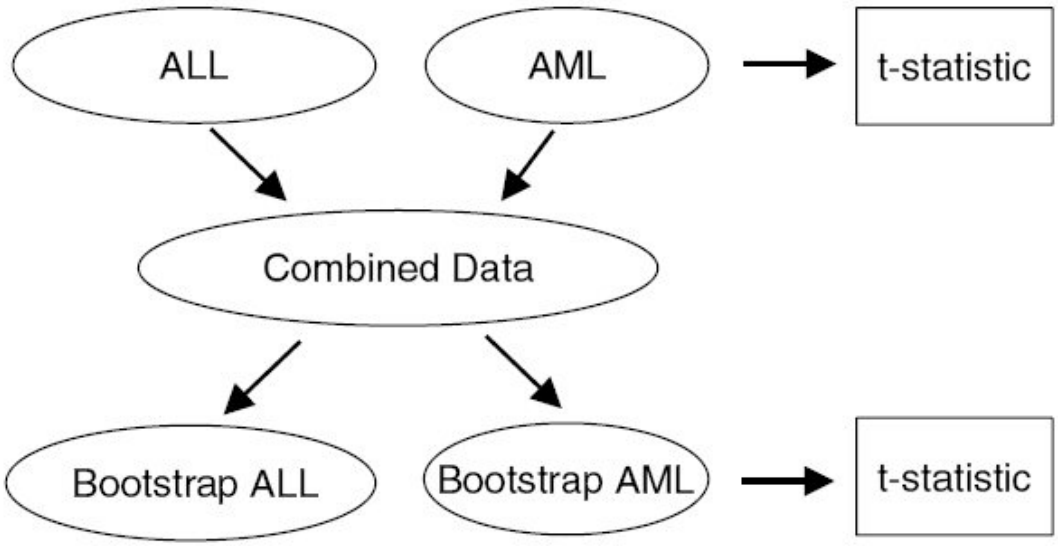


Figura 7.7: La metodologia del bootstrap applicata al data set 7B. Un *t-statistic* è calcolato usando i dati reali, come misura della espressione differenziale del gene, ma un *p*-value non è calcolato direttamente da questo *statistic*. Invece, i dati sono combinati, e quindi sono costruiti i bootstrap data set dai dati combinati. I bootstrap data set hanno anche 27 pazienti ALL ed 11 pazienti AML, ma con ciascun paziente avente una misura scelta casualmente dai 38 valori combinati con i dati originali. Quindi viene calcolato un *t-statistic* per ciascun bootstrap data set per produrre una popolazione di *t-statistic* rappresentante dei dati randomizzati con misure simili ai dati reali. Il *t-statistic* è confrontato con questa distribuzione per generare un *p*-value.

La statistica-t dai dati reali è confrontata con la distribuzione del t-statistic dei dati bootstrap (Figura 7.8). Noi calcoliamo un *p*-value empirico computando la proporzione della statistica bootstrap che ha un valore estremo più grande rispetto al *t*-statistic dei dati reali. Se il t-statistic reale è all'interno della regione a campana della distribuzione, allora è indistinguibile dal t-statistic generato con i dati aleatori. Noi potremmo concludere che il gene non è differenzialmente espresso in modo significativo. Se, dall'altro lato, la statistica dei dati reali si trova verso il margine della distribuzione bootstrap, allora è improbabile che il risultato sperimentale sia casuale, e quindi potremmo concludere che il gene è differenzialmente espresso.

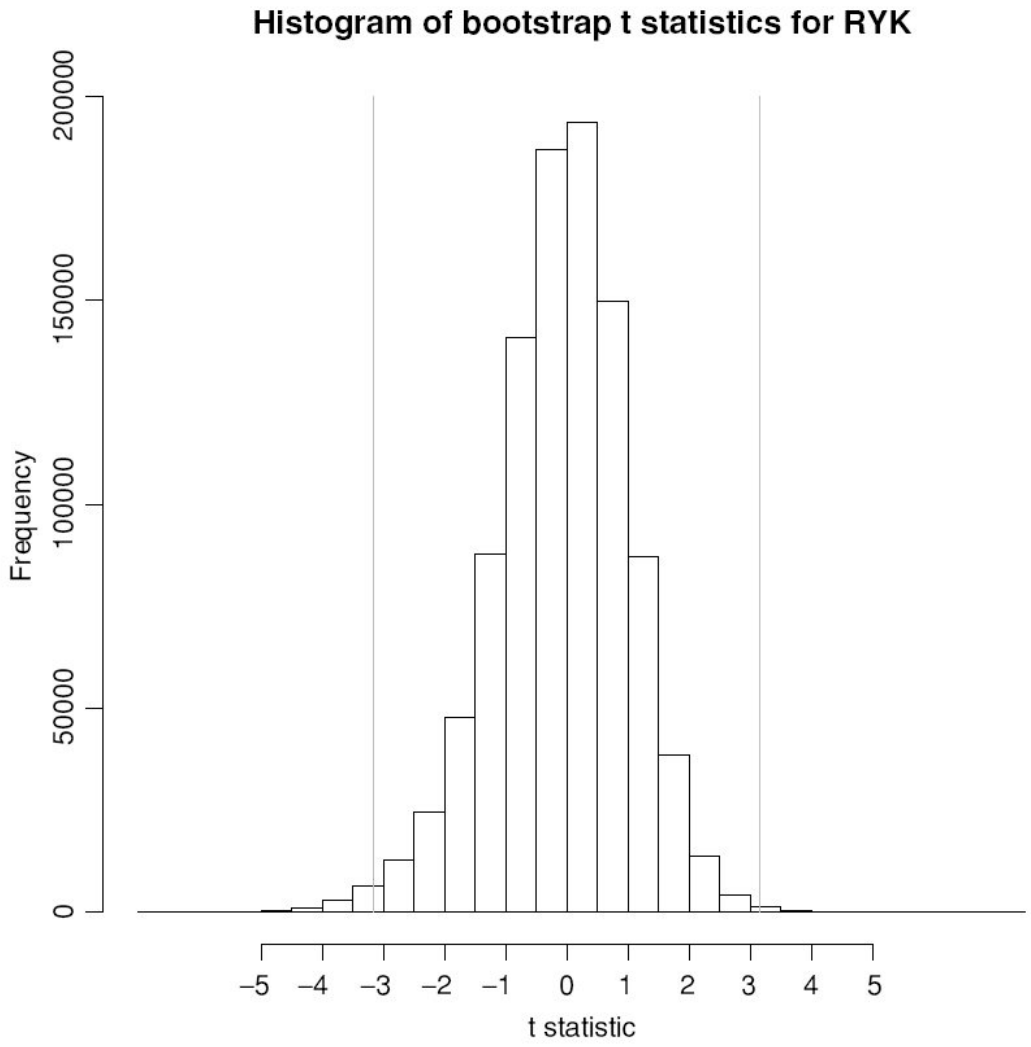


Figura 7.8: Istogramma di risultati del t-statistico con 1.000.000 di ricampionati bootstrap per il gene RYK dai dati della leucemia. Il t-statistic dei dati reali è 3.1596. Noi abbiamo marcato +3.1596 e -3.1596 sull'istogramma per dimostrare che la maggior parte delle statistiche bootstrap sono meno estreme: 9.750 del bootstrap t-statistic giacciono al di fuori di queste linee. Quindi, il *p*-value è proprio al di sotto di 0.0001, e quindi possiamo concludere è espresso differenzialmente in modo significativo ad un livello dell1%.

Esempio 7.10: Bootstrapping RYK dal data set 7B

Consideriamo ancora il gene RYK dal precedente esempio. Vi sono parecchi arrays sui quali il segnale di controllo è di ampiezza più elevata rispetto al segnale del gene, così che il software Affimetrix marchi questo gene come assente. Noi rimpiazziamo queste letture con degli zeri nei dati registrati. Costruiamo il t-statistic per i dati reali, usando l'equazione 7.2. Il t-statistic è di 3.1596, ma noi non utilizziamo il t-distribuzione e per calcolare un *p*-value da questa statistica.

Quindi, noi creiamo i data sets bootstrap, ciascuno dei quali consiste di 27 pazienti ALL ed 11 pazienti AML. Per ciascun paziente, noi scegliamo una misura a caso da 38 valori osservati nella tabella 7.6 ed assegniamo il valore a quel paziente. Per ciascun data set bootstrap costruiamo il t-statistic, usando l'Equazione 7.2, e registriamo il valore. Ripetiamo questo procedimento 1,000,000 di volte per generare la distribuzione bootstrap del t-statistic, (Fig. 7.8). Dei 1,000,000 valori, 9,750 hanno un valore assoluto piú grande di 3,1596, cioè del t-statistic dai dati reali. Quindi il *p*-value bootstrap per il gene RYK è

proprio al di sotto di 0.001, che è significativo al livello dell'1%. Questo contrasta con il risultato del test classico non parametrico. Nel'esempio 7.10, noi abbiamo usato un set di 1,000,000 dati bootstrap per generare una distribuzione. Questo è molto di più di quanto sia di solito necessario. In generale, noi raccomandiamo di sviluppare il bootstrap 10 volte il numero di geni presenti sull'array che stiamo analizzando. Così che con un array di 10,000 geni, raccomandiamo 100,000 dati bootstrap. La ragione di tutto ciò è che il numero di replicati determina la granularità del *p*-value; vedremo nel paragrafo 7.5 che potrebbe essere necessario usare un *p*-value uguale al reciproco del numero di geni sull'array. L'uso di questo numero raccomandato di dati bootstrap permette di fare ciò con ragionevole accuratezza.

Significatività ed Analisi del Microarray

È immediato per un programmatore ragionevolmente esperto scrivere un codice per sviluppare le analisi bootstrap. Tuttavia, esiste un pacchetto software disponibile denominato SAM che sviluppa analisi bootstrap sui dati del microarray. Esso fornisce anche una stima dei falsi positivi usando un metodo che è molto più sofisticato rispetto al metodo descritto nel paragrafo 7.5 SAM è disponibile in Excel in forma di plug-in all'URL indicata alla fine del paragrafo. Esso è molto facile da usare ma può essere lento ove i data sets fossero consistenti. Io non fui capace a sviluppare più di poche centinaia di replicati su due data sets che ho usato negli esempi in questo capitolo usando SAM sul mio PC da tavolo. Tuttavia, esso è utile per piccoli data sets e per lavori esplorativi.

Riassumiamo i vantaggi derivanti dall'uso di SAM oppure altre implementazioni di bootstrap riportate nella tabella 7.7.

TABLE 7.7: Advantages of Different Bootstrap Implementations

SAM	BOOT Package in R	Writing Your Own Program
✓ Easy to use Excel plug-in	✓ Easy to use if you can use R or S+	✓ Maximum speed and efficiency
✓ Can handle many types of data	✓ Faster than SAM	✓ Can run in background on server
✓ Good false discovery rate algorithm		✓ Good for high-throughput analyses

7.5 Molteplicità del testing

Nei paragrafi 7.3 e 7.4 noi abbiamo sviluppato tests statistici su differenti geni ed abbiamo concluso che questi geni possono essere up-regolati e down-regolati sulla base di questi test. Quando noi analizziamo un esperimento di microarray, noi vogliamo applicare questi test in parallelo a molti geni. Nel data set 7A, i dati del cancro alla mammella, vi sono 9,216 geni. Nel data set 7B, i dati della leucemia, vi sono 7,070 geni.

Gli attuali microarrays hanno tipicamente tra 10,000 e 20,000 geni, ed è probabile che in futuro vi saranno array per ciascun splice variante di qualsiasi gene umano con, forse, qualcosa come 300,000 spots. Vi è una importante implicazione nello sviluppo di tests statistici su molti geni in parallelo, che è conosciuta come **molteplicità dei *p*-values**.

Supponiamo di sviluppare un esperimento abbastanza consistente, in cui vi siano microarrays con 10,000 geni. Prendiamo una adeguata fornitura del campione di riferimento, marcato con i fluorocromi Cy3 e Cy5, che co-ibridizza il campione di riferimento a un certo numero di arrays. Prendiamo i dati e sviluppiamo una appropriata normalizzazione per rimuovere la polarizzazione del fluorocromo; quindi sviluppiamo la nostra analisi statistica di scelta su 10,000 geni; questa potrebbe essere un t-test, un Wilcoxon test, oppure un bootstrap test. Poiché ogni campione ibridizzato agli arrays è lo stesso campione di riferimento, noi sappiamo che non vi sono geni differenzialmente espressi: tutte le dissonanze eventualmente misurate nell'espressione sono errori sperimentali. Ma la nostra analisi statistica ci *parlerà di una storia diversa*. Stante alla corretta definizione di un *p*-value, ciascun gene dovrebbe avere una possibilità dell'1% di avere un p-value al di sotto di 0.01, e quindi sarebbe significativo ad un livello dell'1%.

Siccome vi sono 10,000 geni su questo array immaginario, dovremmo aspettarci di trovare 100 geni significativi a questo livello. Similmente, dovremmo aspettarci di trovare 10 geni con un *p*-value inferiore a 0.001, ed 1 gene on p-value inferiore a 0.0001. Ora consideriamo il data set 7A, con 9,216 geni. Benché la terapia non abbia un uguale effetto su ciascun paziente, noi ci aspetteremmo di trovare 92 geni "differenzialmente espressi" con un *p*-value inferiore a 0.01, per il semplice fatto che un elevato numero di geni è stato analizzato. Tutto questo ci fa porre questa importante domanda: Come facciamo a sapere se i geni che sembrano espressi in modo differenziale sono veramente espressi in tal modo, oppure non si tratti di artefatti introdotti per il fatto che noi si stia analizzando un numero elevato di geni? Più concretamente, come dovremmo interpretare il nostro risultato per il gene ACAT2 (Esempio 7.4)? Questo gene è realmente espresso in modo differenziale? Oppure potrebbe essere un falso positivo? Questo è un profondo problema di statistica, che altera molte applicazioni, non soltanto l'analisi del microarray. Descriveremo, quindi, un semplice metodo che può essere usato per stimare la percentuale di geni che sono stati indicati come up-regolati o down-regolati, che hanno una elevata probabilità di essere dei falsi positivi. Un metodo più sofisticato è implementato nel software SAM. Descriveremo anche la correzione di Bonferroni e dimostreremo che questa non è una analisi appropriata per i microarrays.

Stima del rate di falsi positivi

Descriviamo una semplice ma efficiente analisi per stimare il rate dei falsi positivi di un test statistico, e quindi per scegliere un appropriato *p*-value di soglia per geni significativamente espressi in modo differenziale, che fornisce un rate di falsi positivi accettabile. Vi sono 5 passi:

1. Sviluppare una analisi statistica di scelta su ciascun gene che si sta analizzando e memorizzare il *p*-value per ciascun gene.
2. Per un intervallo appropriato di soglie significative, identificare il numero di geni con il *p*-value inferiore a quella soglia.
3. Per le stesse soglie di significatività, calcolare il numero atteso di falsi positivi moltiplicando il *p*-value per il numero di geni che si sta analizzando.
4. Per ciascuna soglia, la percentuale di falsi positivi è il numero aspettato di falsi positivi diviso il numero di geni identificati come espressi a quella soglia.
5. Scegliere una soglia che dia un accettabile rate di falsi positivi.

Esempio 7.11: Molteplicità del *p*-value nel data set 7A

Noi sviluppiamo un test bootstrap su 6350 geni per i quali vi sono dati provenienti da tutti i 20 pazienti dal data set 7A , usando 100,000 bootstrap data sets. Nella tabella 7.8, mostriamo un numero di geni dai dati del cancro alla mammella con differenti intervalli di *p*-value, unitamente al numero di geni che ci si aspetta abbiano quei *p*-values da un testing multiplo. Possiamo vedere dalla tabella 7.8 che usando una tradizionale soglia di significatività dell'1%, che sarebbe stringente nella statistica classica, ci potremmo aspettare di vedere 64 geni significativamente regolati; con i dati reali, noi vediamo 184 geni espressi in modo differenziale. Quindi noi stimiamo che il 35% di questi siano falsi positivi. Dall'altro lato, il numero atteso di falsi positivi con *p*-value inferiore a 0.0001 è 0.6, così è probabile che 14 o 15 esiti su 15 geni osservati e differenzialmente espressi, siano risultati veri positivi. Tuttavia, a questo livello molto stringente, noi perderemo al più 100 geni veri differenzialmente espressi, il nostro rate di falsi negativi è aumentato. Questa illustra che un certo scambio è sempre presente nel controllo dei falsi positivi e dei falsi negativi; una soglia più stringente del *p*-value può portare a meno falsi positivi, ma darà più falsi negativi; una soglia meno stringente sul *p*-value da meno falsi negativi, ma darà più falsi positivi. Il solo modo di migliorare entrambi i rates è di incrementare il numero di individui nello studio; discuteremo ciò nel capitolo 10.

TABLE 7.8: Number and Percentage of False Positives in the Breast Cancer Analysis

For each *p*-value threshold, we count the number of genes observed in the data with *p*-values less than that threshold. This is compared with the expected number of false positives, which is the number of genes being tested multiplied by the *p*-value threshold. The percentage of false positives is the expected number of false positives divided by the observed number of genes. The smaller the threshold used, the fewer false positives, and the better the false positive rate. However, this is at the cost of introducing greater numbers of false negative results.

p-Value Less Than or Equal To	Observed Number of Genes	Expected Number of False Positives	Percentage of False Positives
10^{-2}	184	64	35
10^{-3}	35	6	18
10^{-4}	15	0.6	4
10^{-5}	6	0.06	1

Tabella 7.8: Numero e percentuale di falsi positivi nell'analisi del cancro della mammella. Per ciascun valore di soglia del *p*-value, contiamo il numero di geni osservati nei dati con il p-value inferiore a quella soglia. Questo è confrontato con il numero atteso di falsi positivi, che è il numero di geni che si stanno testando, moltiplicati per la soglia del *p*-value. La percentuale di falsi positivi è il numero aspettato di falsi positivi diviso per il numero osservato di geni. Più è piccola la soglia usata, minore è il numero di falsi positivi, e migliore è il rate di falsi positivi. Tuttavia, questo è il costo che si deve sostenere quando si introducono grandi numeri di falsi positivi.

I *p*-values non aggiustati sono la proporzione dei 100,000 boostrap data sets che hanno un t-statistic più estremo rispetto al t-statistic dei dati reali. Quindi il più piccolo possibile *p*-value è 1/100,000 (oppure 10-5). Dato il numero di geni nel'analisi, i *p*-values corretti di Bonferroni sono tutti troppo grandi per essere significativi, dimostrando che questo metodo non è applicabile alla maggior parte dei dati del microarray.

Esempio 7.12: Scelta dei geni differenzialmente espressi dal data set 7A

In una analisi di esempio del data set 7A, decidemmo che era importante non avere risultati falsi positivi. Pertanto, impostammo una soglia per il *p*-value così che il numero atteso di falsi positivi fosse 1; questa soglia è uguale al reciproco del numero di geni testati, oppure 1/6350 = 1.6 x 10^{-4}. I primi 20 geni in alto derivanti dall'analisi sono elencati nella tabella 7.9. Possiamo vedere dalla tabella che vi sono 16 geni che oltrepassano questa soglia; il rate dei falsi positivi è approssimativamente 1 su 16, circa il 6%.

TABLE 7.9: Significant Genes from the Breast Cancer Data Set

The unadjusted *p*-values are the proportion of the 100,000 bootstrap data sets that had *t*-statistics more extreme than the *t*-statistic from the real data. Thus the smallest possible *p*-value is 1/100,000 (or 10^{-5}). Because of the number of genes in the analysis, the Bonferroni corrected *p*-values are all too large to be significant, illustrating that this method is not applicable to most microarray data.

Accession	Description	*p*-Value	Bonferroni Adjusted *p*-Value
AA598794	connective tissue growth factor	10^{-5}	0.064
N23941	cyclin-dependent kinase inhibitor 1A	10^{-5}	0.064
AA478553	dopachrome tautomerase	10^{-5}	0.064
W96134	v-jun avian sarcoma virus 17 oncogene homolog	10^{-5}	0.064
AA044993	connective tissue growth factor	10^{-5}	0.064
AA040944	v-fos FBJ murine osteosarcoma viral oncogene homolog	10^{-5}	0.064
N95402	copine V	2×10^{-5}	0.13
R12840	v-fos FBJ murine osteosarcoma viral oncogene homolog	3×10^{-5}	0.19
AA442853	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	$4.\times10^{-5}$	0.25
AA418077	GTP-binding protein overexpressed in skeletal muscle	5×10^{-5}	0.32
AA133129	transcription elongation factor B (SIII), polypeptide 3	5×10^{-5}	0.32
AA485377	v-fos FBJ murine osteosarcoma viral oncogene homolog	6×10^{-5}	0.38
AA134757	fibulin 1	6×10^{-5}	0.38
AI831083	dihydropyrimidinase-like 3	7×10^{-5}	0.45
AA004637	ESTs	9×10^{-5}	0.57
No Annotation		1.2×10^{-4}	0.76
AA025939	CD4 antigen (p55)	2×10^{-4}	1.3
H21041	activating transcription factor 3	2.3×10^{-4}	1.5
AA449463	KIAA0220 protein	2.6×10^{-4}	1.7
H05099	KIAA0182 protein	3.8×10^{-4}	2.4

Tabella 7.9: Geni significativi dal data set del cancro della mammella. I *p*-values non aggiustati sono le proporzioni di 100.000 boostratp data set che hanno un *t-statistic* piú estremo rispetto al *t-statistic* dei dati reali. Quindi il valore piú piccolo possibile del *p*-value è 1/000.000 (o 10^{-5}). A causa del numero di geni nell'analisi**,** i valori *p*-values corretti di Bonferroni sono troppo grandi per essere significativi, mostrando che questo metodo non è applicabile alla maggior parte dei dati per microarray.

Correzione di Bonferroni

La colonna finale nella tabella 7.9 da i valori i dei *p*-values, corretti secondo Bonferroni.

La correzione di Bonferroni e un approccio tradizionale per modificare i valori del *p*-values quando si sviluppino molti test statistici in parallelo. Esso è molto simile all'approccio che abbiamo già descritto, ma è più stringente. I *p*-values sono calcolati nel modo normale per mezzo del test statistico, ed i *p*-values sono quindi tutti moltiplicati per il numero di test che vengono sviluppati. Nel caso dell'analisi della espressione del gene, i p-values dovrebbero tutti essere moltiplicati per il numero di geni nell'analisi. Il problema con la correzione di Bonferroni è che essa è di solito troppo stringente per l'analisi dei microarrays. I *p*-value ottenuti sono frequentemente così grandi che nessun gene è considerato differenzialmente espresso. Il prossimo esempio illustra questo fatto.

Esempio 7.13: *p*-Values corretti secondo Bonferroni per il data set 7A

La correzione di Bonferroni è applicata ai dati del cancro alla mammella dell'esempio 7.12. Vi sono 6350 geni testati, cosí che i *p*-values sono moltiplicati per 6350 (tabella 7.8). Benché noi si usi una soglia di significatività molto permissiva del 5%, nemmeno un singolo gene sarebbe indicato come significativo, e quindi questo non sarebbe un buon metodo per l'analisi dei dati.

7.6 ANOVA e Modelli Generali Lineari

Fino a questo punto, noi abbiamo descritto i metodi per analizzare esperimenti molto piccoli per l'espressione differenziale dei geni, in cui i dati sono sia appaiati, con due campioni biologici derivanti dallo stesso individuo, oppure non appaiati, con due gruppi di individui comparati. Man mano che andavamo avanti, i microarrays sono stati usati per sviluppare esperimenti più complessi, in cui ci possono essere piú di due gruppi, oppure in cui è misurata la risposta a più di una variabile. Questi tipi di esperimenti richiedono una analisi più sofisticata come ANOVA e i modelli generali lineari. Noi introdurremo queste idee molto brevemente; vi sono molti libri di statistica di livello intermedo e di livello avanzato che il lettore interessato può consultare in maggiore dettaglio. Questi metodi sono implementati anche in tutti i software statistici, per esempio, SPSS, SAS, S+ e R, che posseggono tutti un documentazione per il loro uso.

Il Metodo ANOVA ad Una Via

Il Data set 7B ha due gruppi di pazienti, e noi eravamo interessati alla comparazione dell'espressione nei due gruppi per identificare differenzialmente i geni espressi.

Supponiamo invece che vi siano tre gruppi di pazienti, e che si sia interessati di identificare i geni che risultino differenzialmente espressi in uno o più gruppi relativamente agli altri. Vi sono due modi in cui potremmo sviluppare l'analisi:

- Semplicemente, potremo applicare per tre volte un *t*-Test per dati non appaiati, a ciascuna coppia di gruppi, e selezionare i geni che sono significativi in uno o più dei *t*-Tests.
- Al contrario, potremmo usare un test statistico che compara tutti e tre i gruppi contemporaneamente e riporti un singolo *p*-value.

Ci sono due problemi con questo metodo. Il primo è la molteplicità: sviluppando tre test, noi aumentiamo la probabilità di vedere una differenza significativa tra due dei gruppi come risultato di errori di misura. Questo problema diventa via via peggiore man mano che i gruppi aumentano: per esempio, con 7 gruppi, ci sarebbero 21 confronti separati. Il secondo problema è che ciascuno di questi confronti non è indipendente dall'altro, così che diventa molto difficile interpretare i risultati. L'approccio intrapreso dagli statistici è il metodo **ANOVA ad una via**. Questo metodo sviluppa una analisi su questo tipo di dati, dove noi stiamo comparando due o piú gruppi, e restituisce un singolo p-value che è significativo se uno o piú gruppi sono differenti uno dall'altro.

Esempio 7.14: Data set 7C e l'analisi ANOVA

Nel data set 7C, i campioni furono presi da quattro gruppi di pazienti affetti da quattro differenti tipi di cancro: neuroblastoma (NB), linfoma non Hodgkin (NHL), rhabdomyosarcoma (RMS) e tumori di Ewing (EWS). Se vogliamo identificare geni che sono espressi un modo differenziale in uno o più di questi quattro gruppi, allora usiamo il metodo ANOVA ad una via. Ciò è molto meglio che sviluppare sei test separati per comparare tutti i gruppi l'uno rispetto all'altro.

ANOVA Multifattore

Nel precedente esempio, vi è soltanto una variabile che condiziona l'espressione del gene: il gruppo a cui l'individuo appartiene, che nei data set 7B o 7C, è il tipo di cancro di cui il paziente è affetto. Supponiamo, tuttavia, che noi stiamo analizzando i dati nei quali i pazienti sono affetti da due tipi di leucemia, e che si conosca anche il sesso di pazienti, che può essere maschio o femmina. In questo caso, l'espressione del gene potrebbe dipendere sia dal tipo di disturbo, che dal sesso del paziente, oppure da entrambi. Modelli ANOVA più generali possono essere costruiti in modo tale che includano due o più fattori, e che ritornino un *p*-value per ciascun fattore separatamente. In questo esempio ci sarebbe un p-value sia che il gene sia differenzialmente espresso per questo tipo di disturbo, sia che non lo sia, ed un altro *p*-value sia/oppure no, che il gene sia differenzialmente espresso in dipendenza del sesso del paziente. Con i metodi ANOVA multi fattore, è possibile che due fattori si comportino in modo concorde in una sorta di maniera additiva o moltiplicativa. Se la risposta ai due fattori è additiva, allora l'effetto di un fattore non influenza l'altro. D'altra parte, supponiamo che ci sia un gene che è differenzialmente espresso nei pazienti maschi del gruppo ALL relativamente al gruppo di pazienti maschi in AML, ma che non é differenzialmente espresso nelle donne. In questo caso, i fattori mostrano un comportamento moltiplicativo, e gli statistici chiamano ciò **interazione** tra i fattori. I pacchetti statistici permettono all'utilizzatore di costruire interazioni con i modelli ANOVA. Frequentemente nel'analisi di microarray, vi possono essere fattori nel'esperimento che non sono di intrinseco interesse scientifico, ma che possono influenzare l'espressione del gene osservata. Per esempio, vi possono essere due o più scienziati che sviluppano la ibridizzazione. In questi casi, noi potremmo desiderare di includere questi fattori entro il modello di analisi statistica; tali fattori sono denominati effetti aleatori e devono essere trattati in modo leggermente diversi.

Modelli Generali Lineari

Le analisi ANOVA sono appropriate la dove i fattori da cui l'espressione del gene dipende siano tutte variabili di categoria, come per esempio il tipo di disturbo oppure il sesso dell'individuo. Alcune volte, tuttavia, si possono avere fattori che sono variabili continue, per esempio, la dose di un composto aggiunto al campione. Se si pensa che l'espressione del gene risponda in modo lineare a questa variabile, allora gli statistici useranno i **modelli generali lineari** per analizzare i dati. I modelli generali lineari possono combinare sia le variabili di categoria che le variabili continue; pertanto, essi sono delle generalizzazioni sia dei modelli ANOVAs che delle regressioni lineari. Così come con i modelli ANOVA, il modello generale lineare restituisce un *p*-value separato per ciascuno dei fattori che vengono testati. È anche possibile includere interazioni nei modelli generali lineari. Supponiamo di avere un esperimento in cui differenti dosi di un composto sono date ai topolini maschi e femmine. Se la risposta alla dose è differente nelle femmine rispetto ai. maschi, allora c'è una interazione tra questi due fattori. Entrambi i modelli ANOVA e modelli lineari generali sono simili al *t*-Test in quanto essi richiedono che la variabilità dei dati sia normalmente distribuita. Tuttavia, è possibile applicare le analisi bootstrap a questi test più sofisticati, che fa sí che essi siano applicabili all'analisi per microarray. Questo approccio è stato seguito in molti articoli che descrivono le applicazioni di ANOVA ai microarray.

Riassunto dei punti chiave

- Le analisi statistiche per i geni espressi in modo differenziale sono sviluppate meglio se si usano test di ipotesi piuttosto che una semplice soglia di rapporto di espressione.
- La struttura dei dati può essere appaiata, non appaiata, oppure più complessa.
- Il tradizionale t-test può non essere appropriato per i dati del microarray poiché questi richiedono che i dati siano normalmente distribuiti.
- I test non-parametrici sono molto robusti al rumore eventualmente presente nel'esperimento, ed i test boostrap sono versioni disponibili, molto potenti, di tali test.
- Il grande numero di geni che si sta testando introduce il problema della molteplicità, e quindi è importante sviluppare una analisi del rate di falsi positivi.
- I dati più complessi possono richiedere l'analisi basata sul modello ANOVA oppure sui modelli generali lineari e possono includere il bootstrapping.

Capitolo 8

Analisi delle Relazioni tra Geni, Tessuti e Trattamenti medici

8.1 Introduzione

I microarray sono una tecnologia genomica. La Genomica è differente dalla Genetica in quanto essa pone l'attenzione non sui geni isolati, ma sul come molti geni lavorano insieme per produrre effetti fenotipici. Nel Capitolo 7 abbiamo visto come i microarrays possono essere usati per lo studio dei geni isolati. Ma gran parte della reale potenzialità dei microarrays consiste nella possibilità dell'uso di essi nello studio delle relazioni tra i geni e nell'identificare i geni di campioni che si comportano in maniera simile o coordinata.

Questo capitolo illustra un certo numero di tecniche di analisi per ricercare e verificare tali relazioni. Facciamo uso di due data set di esempio per esaminare le idee di questo capitolo.

Esempio 8.1: Dati da sporulazione di lievito (data set 8A)

Il lievito gemmante può riprodursi sessualmente producendo cellule aploidi attraverso un processo denominato sporulazione. Il lievito è stato messo in un mezzo sporulante, ed i campioni sono stati presi a sei intervalli di tempo dall'inizio della sporulazione e sono stati, infine, ibridizzati al microarray. Vogliamo identificare i gruppi di geni che si comportano in modo coordinato durante l'evoluzione di questa serie temporale[1].

Esempio 8.2: Sottotipi di linfoma diffuso a cellule B (data set 8B)

I campioni sono stati presi da 39 pazienti affetti da linfomi diffusi a grandi cellule B ed ibridizzati ai microarrays. Vogliamo identificare i geni che sono co-regolati in questa malattia. Siamo interessati anche alla possibilità che vi siano gruppi di pazienti con profili di espressione genica similare[2].

Questo capitolo discute i metodi che possono essere usati per rispondere a questo genere di domande; esso è organizzato nei seguenti cinque paragrafi:

1. *Paragrafo 8.2*: Similarità Campione nei profili di geni o campioni, ponendo attenzione ai differenti metodi per quantificare la similarità oppure la non-similarità nei profili di espressione genica. Mostreremo come differenti metodi possano dare differenti risultati e, dunque, evidenzieremo la necessità di valutare con cura la scelta da effettuare riguardo al metodo da usare.
2. *Paragrafo 8.3*: Riduzione della dimensionalità, descrive due metodi per ridurre la dimensionalità dei dati: analisi delle componenti principali e scaling multidimensionale. Uno dei problemi dell'analisi dei dati dei microarrays consiste

[1] I dati sono stati presi dal lavoro di Chu ed altri (1998). Un riferimento completo é dato alla fine del capitolo, ed i dati sono disponibili allo Stanford Microarray Database

[2] I dati furono presi dal lavoro di AlizadeH ed altri (2000),). Un riferimento complete é dato alla fine del capitolo, ed i dati sono disponibili allo Stanford Microarray Database

nel fatto che è molto difficile per il cervello umano concettualizzare il grande numero di geni e campioni tipicamente coinvolti. Per esempio, nel data set 8B, ci sono 40 campioni ibridizzati ai microarray con circa 18000 geni. Questi metodi possono ridurre i dati di 2 o 3 ordini di grandezza, permettendo "all'operatore umano" di visualizzarli facilmente; vi sono anche utili tools per usare i metodi di classificazione descritti nel capitolo 9

3. *Paragrafo 8.4:* Clustering gerarchico, introduce il metodo comunemente usato per l'identificazione dei gruppi di geni (oppure di tessuti) strettamente correlati. Il clustering gerarchico è un metodo che collega in successione geni oppure campioni che abbiano una struttura similare per formare una struttura ad albero, molto simile ad un albero filogenetico. Discutiamo differenti versioni dell'algoritmo e mostriamo come vi possano essere differenti risultati con gli stessi dati.
4. *Paragrafo 8.5:* L'Affidabilità e la Robustezza del Clustering Gerarchico descrive i metodi per validare il clustering gerarchico. Poniamo particolare attenzione sul bootstrapping e sulla costruzione dell'albero di consenso, che possono essere usati sia per validare i cluster, sia per impostare una misura numerica di confidenza su di essi.
5. *Paragrafo 8.6:* Metodo Machine Learning (n.d.t.: una Learning Machine è un automa a stati finiti che fornisce un risultato basato sullo stato presente delle variabili, ma anche, e soprattutto, sulla memoria del passato) per l'analisi del cluster, descrive due ulteriori metodi per il clustering dei dati, entrambi messi a punto nella comunità di scienziati appartenenti alla Machine Learning. K-means clustering è un metodo di clustering non gerarchico che richiede all'analista di fornire preventivamente il numero dei cluster, e quindi alloca i geni ed i campioni nei vari cluster in modo appropriato. La mappa di auto-organizzazione è un metodo appartenente alla stessa classe; esso alloca i geni o i campioni ad un numero predefinito di cluster che sono correlati l'un l'altro su una griglia spaziale. Entrambi i metodi sono implementati in un grande numero di pacchetti software per l'analisi dell'espressione dei geni.

8.2 Similarità del Gene oppure Profili Campione

Quando noi usiamo un microarray come strumento di ricerca genomica, noi vogliamo identificare i geni o i campioni che abbiano profili di espressione similari. Una volta che abbiamo prodotto grafici o tabulati di tali profili, molte persone potrebbero giá avere una buona intuizione di cosa questo significhi. Quando sviluppiamo un analisi computazionale sui dati, è necessario che si sia in grado di trasformare queste intuizioni in misure quantitative che possano essere elaborate via software, e che riflettano questa intuizione in un modo affidabile e robusto.

In questo capitolo abbiamo volutamente scelto due data set alquanto differenti: il data set 8A è una serie temporale, e il data set 8B riguarda i campioni di un gruppo di pazienti. Descriveremo un certo numero di misure di similarità tra i profili e mostreremo esempi di queste misure applicate ad essi. Quindi, mostreremo come due profili possano essere similari nel'ambito di una misura, ma differenti in un altro ambito di misura - pertanto, puntualizzeremo l'importanza di scegliere una misura appropriata di similarità per le tecniche di analisi che descriveremo nel seguito di questo capitolo.

Vi sono due modi per esaminare i dati del microarray: sia che si sia interessati alla similarità dei geni, sia che si sia interessati alla similarità dei campioni.
Nel primo caso, ciascun gene è misurato per mezzo dei campioni; nel secondo caso, ciascun campione è misurato attraverso i geni (Figura 8.1). Dal punto di vista di una

prospettiva scientifica, vi sono differenti analisi. Dal punto di vista dei metodi di analisi dei dati in questo capitolo, invece, queste sono essenzialmente simili. Pertanto, nel corso di questo capitolo, noi ci riferiremo ai *profili genici*, sapendo che questi possono essere sia profili di geni che di campioni.

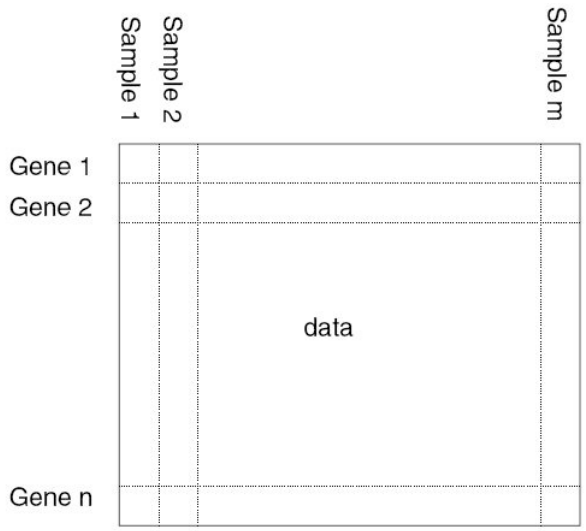


Figura 8.1: Due modi di guardare agli stessi dati. La matrice dei dati di *m* campioni ed *n* geni può essere analizzata in due modi. Sia che si guardi alle relazioni tra i geni, usando l'espressione di ciascuno dei campioni come misure dei geni, sia che si guardi per le relazioni tra campioni, usando l'espressione in ciascun gene come misure dei campioni.

Caratteristiche delle Misure di Distanza

È consuetudine comune descrivere la similarità tra due profili in termini di distanza fra di essi in uno spazio a molte dimensioni dell'espressione dei geni, o delle misure dei campioni. Prima di descrivere le specifiche misure di distanza, richiamiamo alcune proprietà teoriche che una misura di similarità (o di non similarità) tra due geni dovrebbe avere. Queste possono apparire ovvie, ma esse sono molto importanti se si usa una misura di similarità come efficace descrittore dei dati.

- La distanza tra due qualsiasi profili deve essere maggiore di zero o uguale a zero - la distanza non può essere negativa.
- La distanza tra un profilo e se stesso deve essere zero.
- Inversamente, se la distanza tra i due profili è zero, allora i profili devono essere identici.
- La distanza tra il profilo A ed il profilo B deve essere uguale a quella tra il profilo B e il profilo A.
- La distanza tra il profilo A ed il profilo C deve essere minore o uguale alla somma delle distanze tra i profili A e B ed i profili B e C.

Le prime quattro regole sono abbastanza intuitive; l'ultima regola consiste in quella che è conosciuta come *triangolo della disuguaglianza* (Figura 8.2)

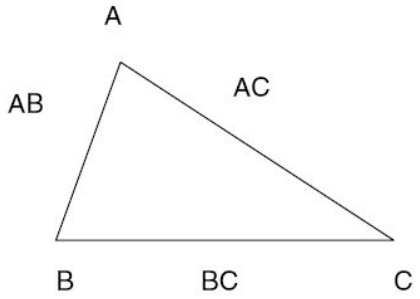


Figura 8.2: Il triangolo della disuguaglianza. Il triangolo ABC ha le distanze AB, AC e BC. Queste distanze soddisfano le seguenti equazioni:
AB ≤ AC + BC
AC ≤ AB + BC
BC ≤ AB + BC
Queste equazioni significano che la lunghezza di ciascun lato è inferiore o uguale alla somma delle lunghezze degli altri due lati. Se le lunghezze sono uguali, allora tutti e tre i punti sono sulla stessa linea retta ed il triangolo è completamente piatto.

Coefficiente di Correlazione

La prima misura di similarità che andiamo a descrivere è il coefficiente di correlazione. Vi è un concetto statistico che esprime in modo quantitativo il livello di relazione tra due set di misure (Figura 8.3).

Se noi indichiamo i due set di misure con la notazione x(i) ed y(i), dove i è l'indice che va da 1 ad n, allora il coefficiente di correlazione r è dato dalla seguente formula:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

Equazione 8.1

Il coefficiente di correlazione assume un valore compreso tra -1 e +1. Un valore di -1 rappresenta una forte correlazione negativa: quando una variabile è alta, l'atra è bassa. Un valore di +1 rappresenta una forte correlazione positiva: quando una variabile è alta, anche l'atra è alta. Un valore pari a zero rappresenta variabili non correlate.

È pratica comune - prima di calcolare il coefficiente di correlazione - centrare i profili di espressione del gene (Paragrafo 5.4) per essere sicuri che essi abbiano la media uguale a zero e la deviazione standard uguale ad uno. Quando il profilo è stato centrato, il coefficiente di correlazione è dato dalla seguente semplice formula[3]:

$$r = \sum_{i=1}^{n} x_i y_i$$

Equazione 8.2

[3] Questa formula é il prodotto scalare di due set di misure (x_i) e (y_i) quando si pensi ai dati come un vettore nello spazio n dimensionale. Il prodotto scalare misura l'angolo tra i due vettori, e così il coefficiente di correlazione ha una interpretazione geometrica: vettori paralleli hanno coefficiente di correlazione pari a -1 o +1, mentre vettori ortogonali hanno coefficiente di correlazione pari a 0. Per convertire tutto ciò in distanza, é necessario applicare l'equazione 8.3 o 8.4 in modo che i vettori paralleli abbiano distanza 0 e i vettori ortogonali abbiano distanza 1

Il centraggio dei dati è un metodo eccellente per dati simili a quelli del data set 8B, dove stiamo osservando l'espressione dei geni in pazienti relativi al campione di riferimento. Tuttavia, quando stiamo analizzando i dati della serie temporale, come i dati del data set 8A, dove i dati sono relativi all'espressione del gene al tempo 0, il centraggio può portare alla perdita della naturale nozione di geni che sono up-regolati o down-regolati nelle serie temporali. Questo può essere uno svantaggio. Il coefficiente di correlazione è una misura della similarità e necessita di essere convertito in una misura di distanza con le proprietà elencate precedentemente. Vi sono un certo numero di formule che possono essere usate per raggiungere lo scopo, incluse le seguenti:

$$d(\mathrm{X},\mathrm{Y}) = 1 - \mathrm{abs}(r(\mathrm{X},\mathrm{Y}))$$

Equazione 8.3

$$d(\mathrm{X},\mathrm{Y}) = 1 - r(\mathrm{X},\mathrm{Y})^2$$

Equazione 8.4

L'equazione 8.3 stabilisce che la distanza tra i profili X ed Y è uguale ad 1 meno il valore assoluto del coefficiente di correlazione tra X ed Y. Cosí se X ed Y sono totalmente negativi o positivi (r = -1 o r = +1), rispettivamente), la distanza tra di essi è 0, e se essi sono totalmente scorrelati, la distanza tra di essi è 1. L'equazione 8.4 è molto simile, ma viene sottratto il quadrato del coefficiente di correlazione invece del valore assoluto.

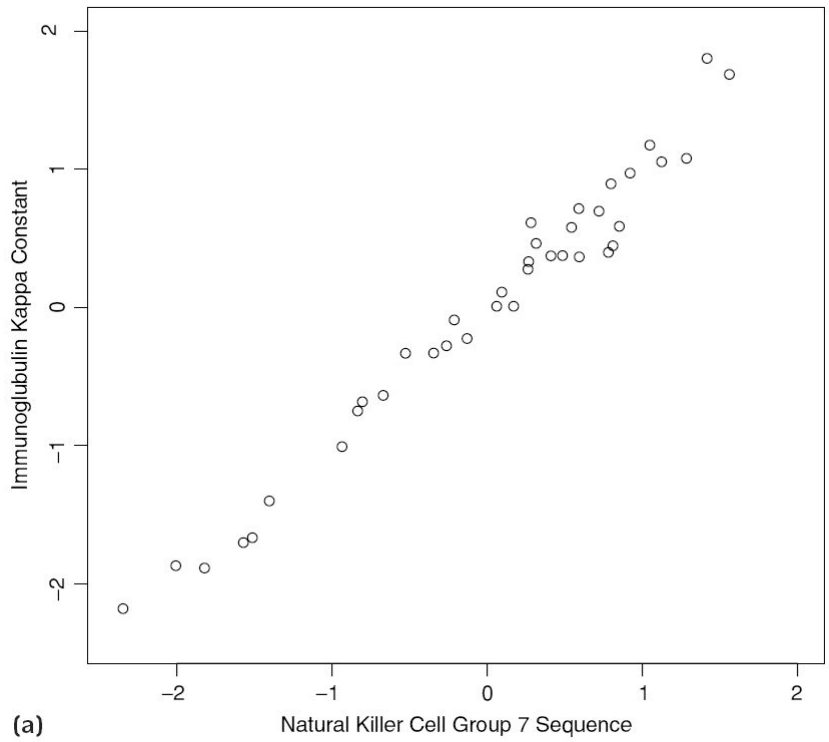
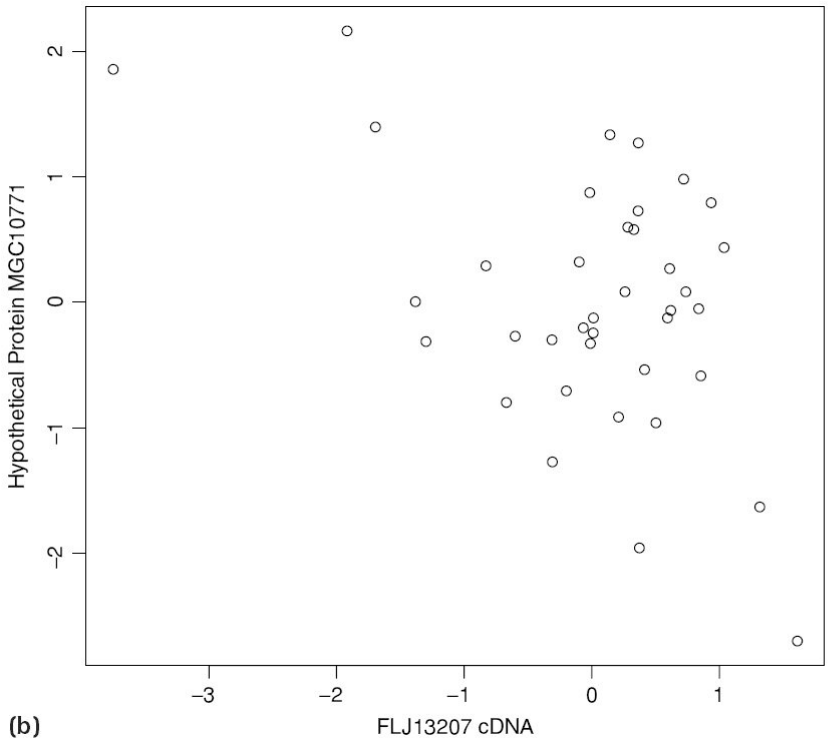
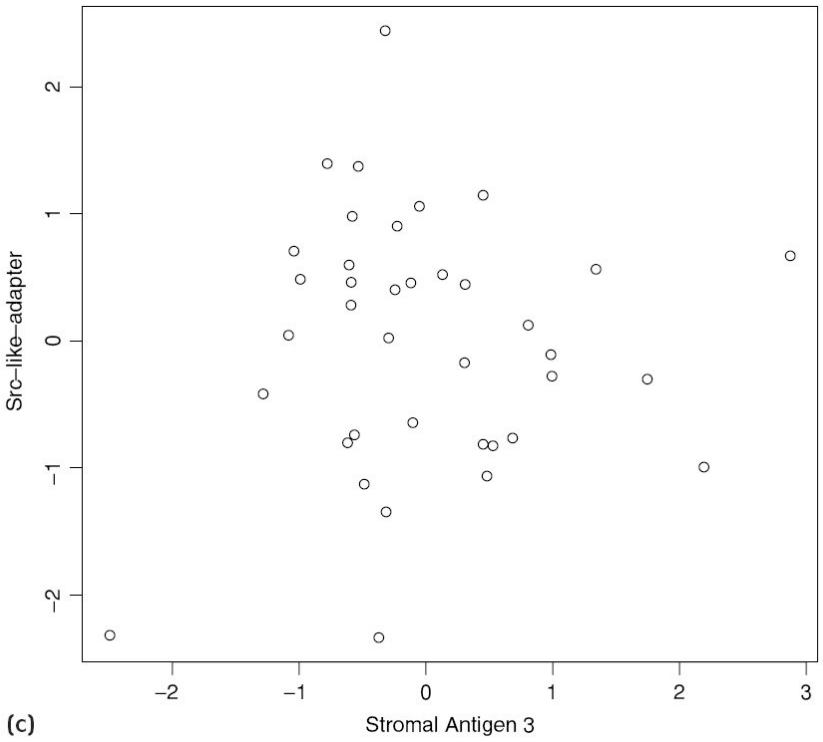


Figura 8.3: Variabili correlate e scorrelate. Queste tre figure mostrano esempi di variabili correlate e scorrelate usando i dati del data set 8B. Per ciascun gene, noi abbiamo calcolato il rapporto logaritmico dell'intensità di quel gene relativo al campione di riferimento. Ciascun gene è stato centrato, così che i valori di ciascun gene abbiano media nulla e deviazione standard uguale ad 1. Ciascun punto del grafico rappresenta uno dei 38 pazienti, dove il valore sull'asse x è il rapporto logaritmico del gene ed il valore sull'asse y è il rapporto logaritmico di un altro gene. **(a)** I geni Cell Natural Killer del Gruppo 7 (NGK7) e la Immunoglobulina Kappa Constant (IGKC) sono positivamente correlati in modo molto marcato (r = 0.97).

I 38 punti giacciono su una linea retta dal basso a sinistra all'alto a destra. **(b)** I due geni non annotati FLJ13207 e MGC10771 sono negativamente correlati in modo debole (r = -0.47). I 38 punti giacciono approssimativamente su una linea dall'alto a sinistra al basso a destra – ma questo è principalmente l'effetto delle tre misure nella regione in alto a sinistra del diagramma, e delle tre misure in basso a sinistra. La maggior parte dei punti si trovano entro una nuvola nel mezzo. **(c)** I geni Stromal Antigen 3 (STAG3) e l'Scr-like-adapter non sono correlati (r = 0.054). I 38 punti giacciono in una curva intorno allo zero, dove non si può discernere in modo chiaro la tendenza.



(b)



(c)

Correlazione di Spearman

Uno dei problemi derivanti dall'uso del coefficiente di correlazione standard appena definito, è che esso è suscettibile di deviazione causata dai dati fuori range: un singolo dato può far sembrare due geni correlati quando tutti gli altri punti suggeriscono che non lo sono. La correlazione di Spearman è una misura non parametrica della correlazione ed è molto robusta rispetto ai dati estremi, e quindi - grazie a questa proprietà - è spesso piuttosto appropriata per l'analisi dei microarrays.

Esempio 8.3: Serie temporali falsamente correlate

La figura 8.4 mostra due geni presi dalla serie temporale del data set 8A, ENB1 e NPR2, che sembrano essere ragionevolmente correlati, con un coefficiente di correlazione (correlazione di Pearson) di 0.63.

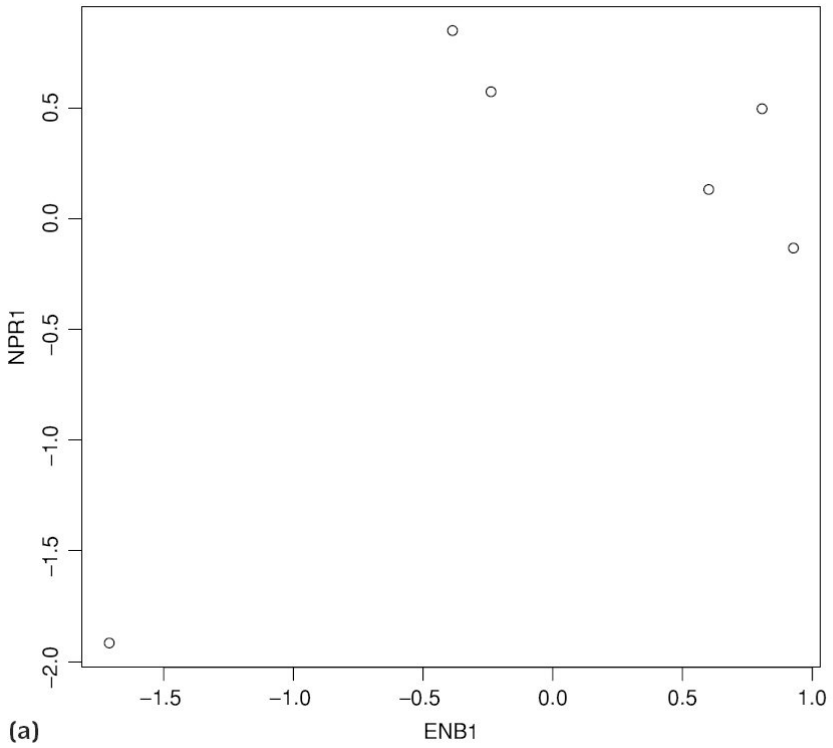
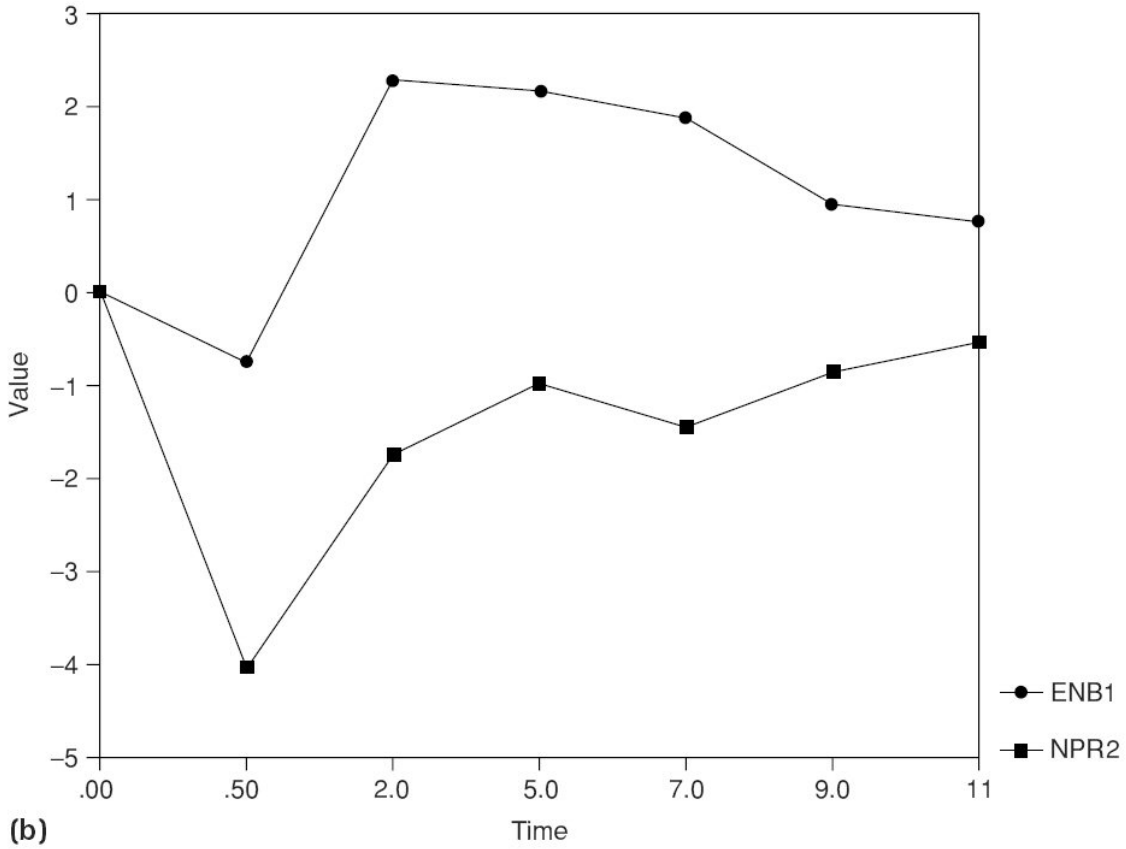
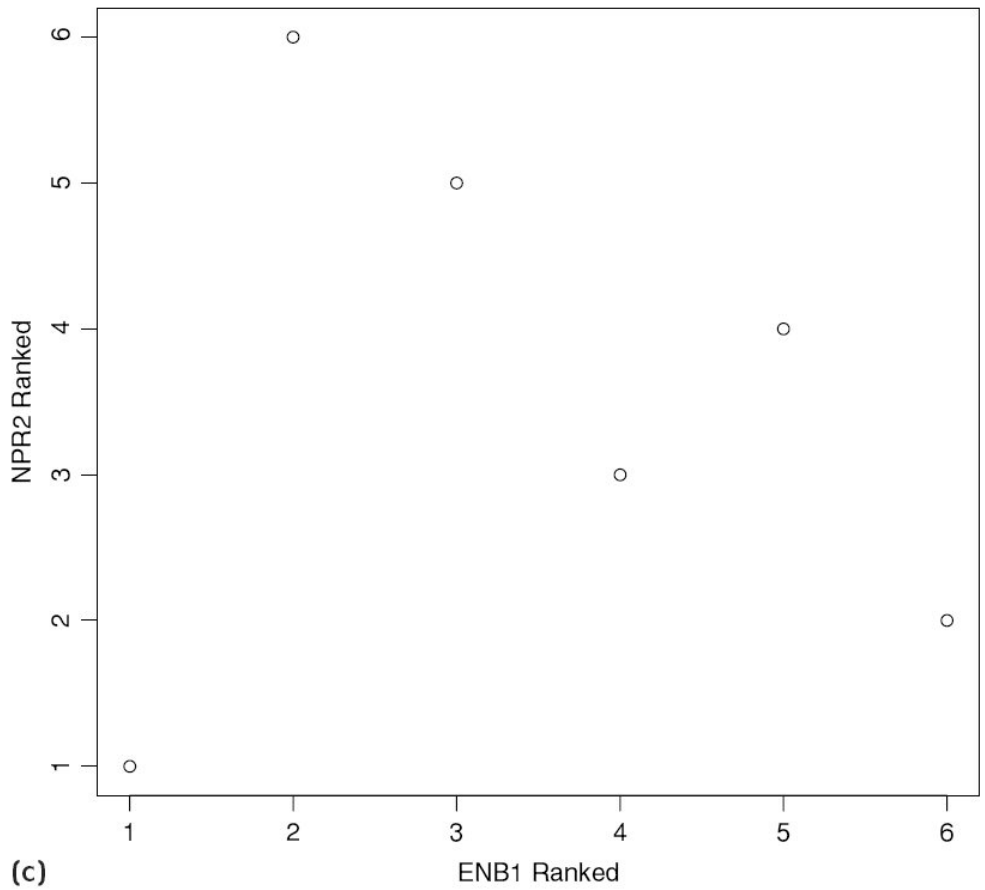


Figura 8.4: Serie temporali falsamente correlate. Mostriamo i dati provenienti da due geni del data set 8A: ENB1 e NPR1. **(a)** Diagramma di correlazione. Per questo diagramma, i valori di ciascuna serie temporale sono stati centrati cosí che essi abbiano media 0 e deviazione standard uguale a 1. Il coefficiente di correlazione è 0.63, un valore mediano positivo della correlazione. La correlazione positiva è un risultato di un singolo dato fuori range al basso a sinistra della figura. I dati degli altri punti non sono positivamente correlati. **(b)** Diagramma di serie temporale. Su questo grafico, abbiamo riportato il rapporto logaritmico (in base 2) del segnale di ciascun punto relativo al segnale al tempo zero. I due grafici mostrano che non vi è affatto correlazione. Il dato fuori range che risulta fortemente correlato si trova a 30 minuti, dove entrambi i geni sono down-regolati. D'Altra parte il comportamento dei due geni dopo la down-regolazione iniziale è completamente differente, con ENB1 che mostra una diminuzione nel'espressione e NPR2 che mostra un incremento nel'espressione. **(c)** Diagramma Ranked. Ai valori delle serie temporali è stato attribuito un rango così che il valore più basso in ciascuna serie abbia rango 1, ed il più alto abbia rango 6, e quindi riportato in grafico. Qui non sembra esserci correlazione tra i punti delle serie temporali dopo l'operazione di attribuzione del rango.


3
2
1
0
−1
−2
−3
−4
−5
Value
.00
.50
2.0
5.0
7.0
9.0
11
Time
ENB1
NPR2

(b)


6
5
4
3
2
1
NPR2 Ranked
1
2
3
4
5
6
ENB1 Ranked

(c)

TABLE 8.1: Spearman Correlation

Time	ENB1 Ratio	NPR2 Ratio	ENB1 Rank	NPR2 Rank
0.5	−0.76359	−4.05957	1	1
2	2.276659	−1.7788	6	2
5	2.137332	−0.97433	5	4
7	1.900334	−1.44114	4	3
9	0.932457	−0.87574	3	5
11	0.761866	−0.52328	2	6

Nota: I valori non centrati per ciascuno dei sei punti temporali per i geni ENB1 e NPR2 sono mostrati nelle colonne 2 e 3. I Ranghi sono calcolati ordinando i valori di ciascun gene ed assegnando il valore 1 per la misura piú bassa, ed il valore 6 per la misura più alta. L'equazione 8.1 viene quindi applicata per calcolare il coefficiente di correlazione: in questo caso la correlazione è -0.09 (correlato negativamente in modo marginale). Di converso, la correlazione di Pearson è 0.63 (correlato positivamente in modo marcato).

Tuttavia, noi possiamo vedere dalla figura 8.4a che la ragione per questa correlazione è il dato fuori range nell'angolo inferiore sinistro; se questo fosse rimosso, i punti non apparirebbero positivamente correlati. La situazione è riflessa nei diagrammi delle serie temporali (Figura 8.4b): i geni non appaiono essere correlati.

I dati fuori range costituiscono un problema serio e purtroppo molto comune nei dati dei microarrays; questo è in parte dovuto ai dati che possono essere rumorosi, ed in parte è dovuto al grande numero di geni che si stanno studiando. La correlazione di Spearman fornisce una misura della correlazione che ha la peculiarità di essere molto robusta in presenza di dati piuttosto estremi .

La correlazione di Spearman lavora in maniera simile ai test non parametrici descritti al Cap. 7. Le misure vere oppure i rapporti logaritmici sono sostituiti dai ranghi: 1 per il valore più basso, 2 per il secondo valore più basso, e così via[4]. L'equazione 8.1 viene quindi applicata ai dati ranked, che producono un coefficiente di correlazione che varia da -1 a 1. Per usare la correlazione di Spearman come una misura di distanza, applichiamo l'equazione 8.3 o 8.4 così che le variabili scorrelate abbiano distanza 0.

Esempio 8.4: Correlazione di Spearman di ENB1 e NPR2

La procedura per calcolare la correlazione di Spearman si può vedere in Tabella 8.1. La correlazione di Spearman di questi geni è -0.09, in confronto con la correlazione di Pearson che è di 0.63. Questo è un risultato molto diverso: piuttosto che essere positivamente correlate, le serie temporali sono probabilmente non correlate (Figura 8.4c).

In generale, la correlazione di Spearman è una misura più robusta rispetto alla correlazione di Pearson, e pertanto può essere più appropriata per i dati del microarray, in modo particolare se nei dati ci dovesse essere rumore. Tuttavia, così come con la centratura dei dati per la correlazione di Pearson, la direzione di regolazione dei geni viene perduta durante il processo di assegnazione dei ranghi.

[4] I punti dei dati vincolati sono presi dalla media dei ranks vincolati. Per esempio, se due piccoli insiemi di data points fossero vincolati, entrambi i punti dovrebbero dare un rank di 1.5.

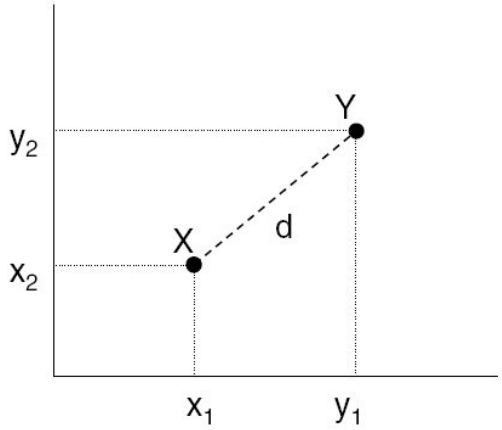


Figura 8.5: Distanza Euclidea. La distanza tra due punti X ed Y nello spazio a due dimensioni, con coordinate (x_1, y_1) e (x_2, y_2), è data dal teorema di Pitagora:

$$d(\mathrm{X,Y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Questo concetto si estende naturalmente anche agli spazi con dimensione superiore (Equazione 8.5)

Questo non è un problema con i dati dei pazienti (e.g., il data set 8B), ma può dare problemi con le serie temporali (e.g., data set 8A). Mostreremo un esempio di questo più avanti.

Distanza Euclidea

La distanza Euclidea è molto diversa dalla correlazione intesa come misura della relazione dei profili di espressione dei geni. La distanza Euclidea è una estensione della distanza che usiamo nel mondo reale: la distanza in linea retta tra punti nello spazio a due o tre dimensioni.

Nel caso di due dimensioni, la distanza tra due punti è calcolata usando il teorema di Pitagora (Figura 8.5). Quando usiamo la distanza Euclidea con i profili di espressione dei geni, estendiamo la stessa idea ad un numero maggiore di dimensioni. Usando la stessa notazione così come abbiamo fatto prima, la distanza Euclidea tra due profili X ed Y è data alla seguente equazione:

$$d(\mathrm{X,Y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Esempio 8.5: La Distanza Euclidea non è scala-invariante

Uno dei problemi chiave con la distanza Euclidea è che essa non è scala invariante: due profili di espressione del gene con la stessa forma ma con differenti ampiezze appariranno molto distanti fra loro. I geni BUR6 e IDH1 dal data set 8A hanno profili simili per ciò che riguarda la up-regolazione, raggiungendo il picco di espressione del gene dopo 7 ore (Figura 8.6).

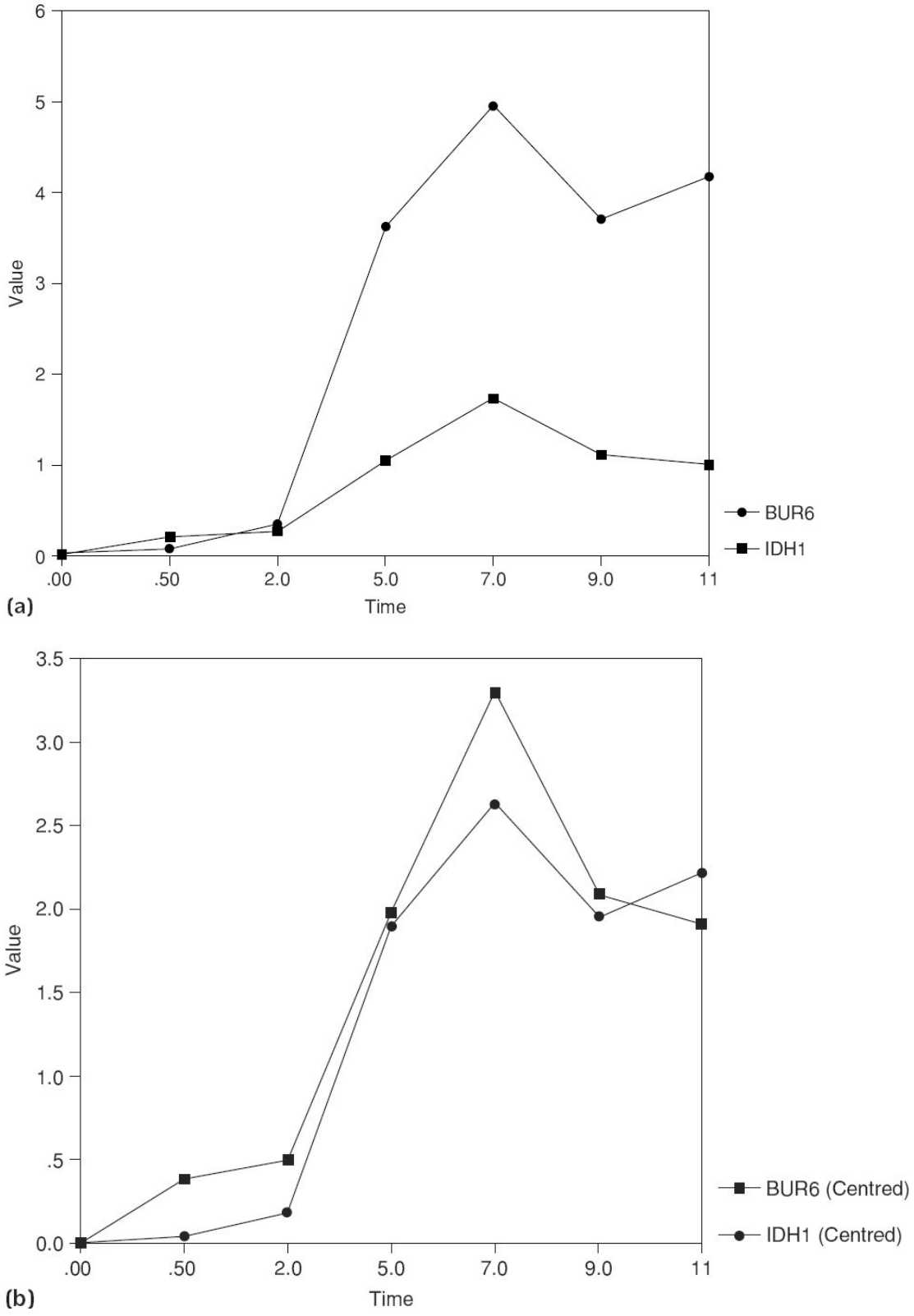


Figura 8.6: La distanza Euclidea non è scala invariante. sono mostrati i due geni bur6 e IDH1 del data set 8B. Entrambi i geni hanno profili simili e sono up-regolati, raggiungendo il massimo dell'espressione del gene dopo 7 ore. **(a)** Dati del rapporto logaritmico. Benché i profili abbiano una forma molto simile, BUR6 è considerevolmente più up-regolato, raggiungendo un massimo di 5 (una up-regolazione di circa 30 volte) mentre IDH1 raggiunge un massimo di circa 2 (una up-regolazione di circa 4 volte). La distanza Euclidea è 5.8 (un valore elevato). **(b)** I dati sono stati scalati dividendo ciascun profilo per la deviazioni standard dei valori assoluti dei punti temporali. Questo metodo preserva la direzione di regolazione dei geni relativi al tempo zero. La distanza Euclidea è di 0.88 (un valore molto più piccolo).

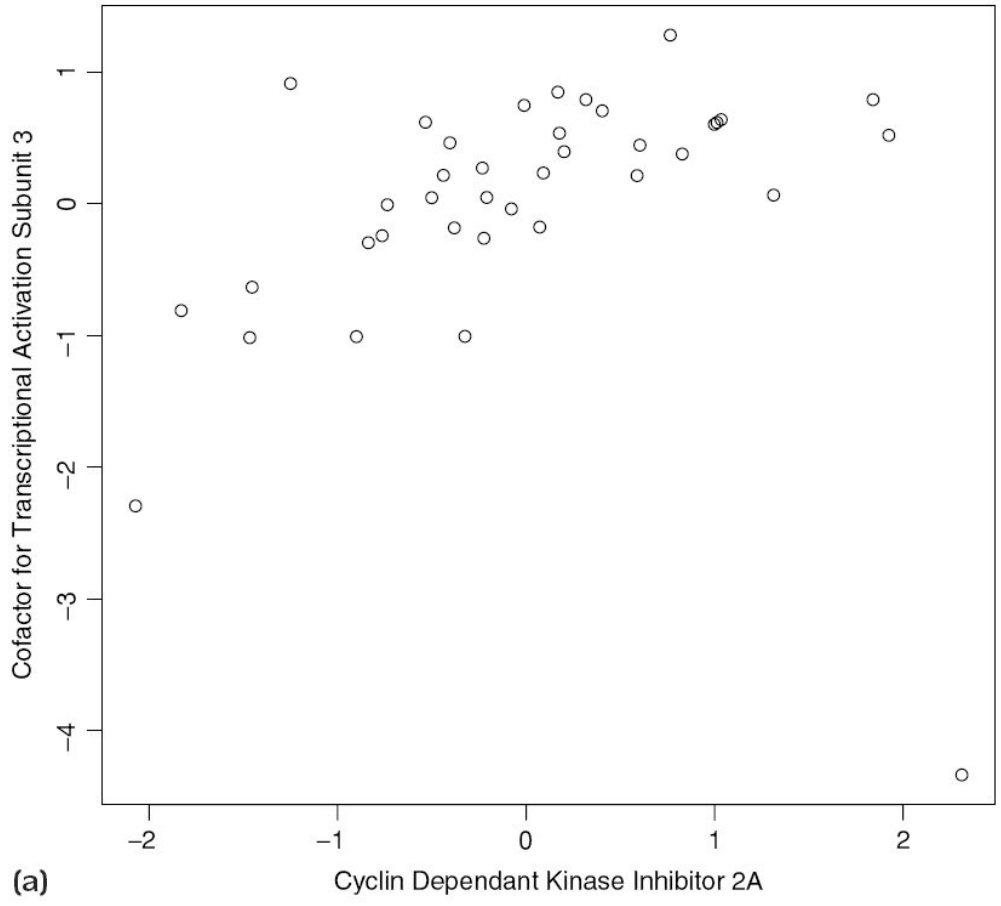
Tuttavia, BUR6 é più up-regolato rispetto a IDH1. La distanza Euclidea tra questi geni è molto grande, benché i profili siano molto simili nella forma.

Questo problema può essere risolto centrando il profilo. Quando il profilo è stato centrato, la distanza Euclidea è molto piccola (Figura 8.6b).

Esempio 8.6: la Distanza Euclidea e la Correlazione possono dare differenti risultati

Nella Figura 8.7 mostriamo due esempi di profili del gene che appaiono simili dal punto di vista della correlazione, ma diversi dal punto di vista della distanza Euclidea. Nel primo esempio consideriamo i dati del data set 7B, dove noi siamo interessati ad identificare quei geni con comportamento simile in differenti pazienti. In questo caso, la correlazione di Spearman fornisce un risultato più robusto: i dati sono positivamente correlati, ma un singolo elemento fuori posto nella direzione opposta alla correlazione è risultato con una grande distanza Euclidea.

Il secondo esempio è la serie temporale dei dati del data set 8A: i due geni hanno simile forma dei profili, ma uno è up-regolato mentre l'altro è down-regolato. Essi sembrano essere correlati, benché l'espressione sia molto differente; la distanza Euclidea è grande, fatto questo che riflette la differenza tra i patterns. In questo caso, la distanza Euclidea è probabilmente una misura più realistica.



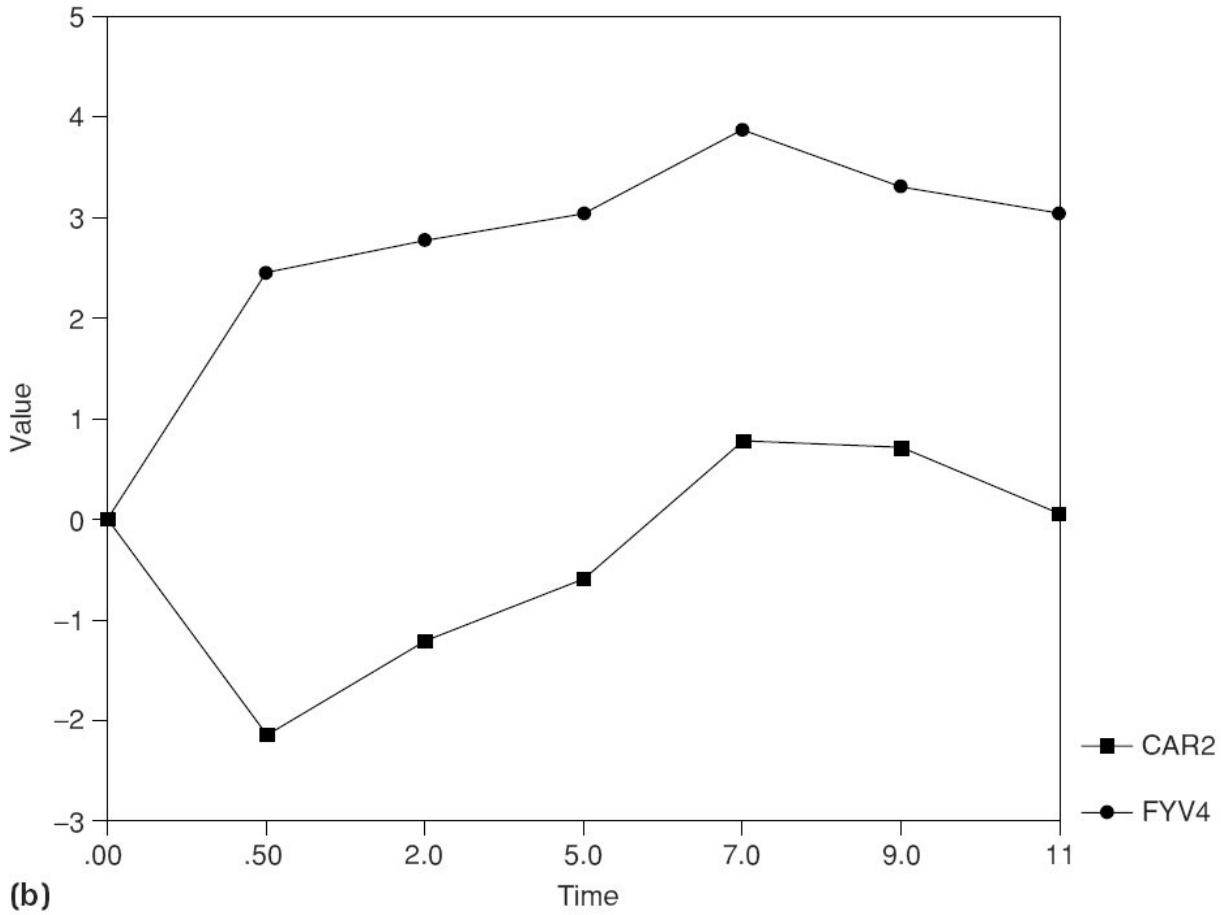


Figura 8.7: Differenza tra distanza Euclidea e correlazione di Spearman. (a) I dati del data set 8B dove la distanza Euclidea è grande (7.9), ma dove la correlazione di Spearman è anche grande (r = 0.79 oppure la distanza = 0.21). Il diagramma mostra i dati centrati (che hanno media 0 e deviazione standard 1) per i due geni Cyclin Dependant Kinase Inhibitor 2A (CDKN2A) e Cofactor for Transcriptional Activation Subunit 3. I due geni sono positivamente correlati, ma vi è un singolo elemento fuori posto che è nella direzione opposta al trend principale (angolo destro inferiore della figura). Questo elemento fuori posto fa in modo che ci sia un elevato valore della distanza Euclidea, mentre la correlazione di Spearman rivela che c'è correlazione. In questo caso, la correlazione di Spearman è probabilmente la migliore misura di similarità.
(b) I dati del data set 8A dove la distanza Euclidea è elevata (8.5) ma la correlazione di Spearman è molto forte (r = 0.91 oppure la distanza = 0.09) per i due geni CAR2 e FYV4. Il diagramma mostra i dati del rapporto logaritmico (base 2); la distanza Euclidea è stata calcolata sui dati che sono stati scalati per mezzo della deviazione standard dei valori assoluti dei rapporti. In questo caso, CAR2 è circa è circa 4 volte down-regolato dopo 30 minuti, mentre FYV4 è circa 5 volte up-regolato allo stesso valore del tempo. Dopo 30 minuti, i geni hanno la stessa forma, ma FYV4 rimane up-regolato, mentre CAR2 passa da down-regolato ad up-regolato. In questo caso la distanza Euclidea è probabilmente la migliore misura di similarità.

Da questi esempi, possiamo vedere che quando si scelga una misura di distanza da usare per ulteriori analisi, come ad esempio l'analisi del cluster che discuteremo più tardi nel corso del capitolo, non vi è una risposta su quale sia la migliore misura. Differenti misure hanno differenti punti di forza e di debolezza, e possono essere combinate con differenti scaling dei dati per produrre differenti risultati (Tabella 8.2).

TABLE 8.2: Strengths and Weaknesses of Different Distance Measures

Pearson Correlation	Spearman Correlation	Euclidean Distance
✓ Powerful	✓ Robust to outliers	✓ Geometric interpretation
✓ Spots positive and negative correlations	✓ Spots positive and negative correlations	✓ Can retain up- or down-regulation information with appropriate scaling
✓ Scale invariant on centred data	✓ Completely scale invariant: no scaling or centering required	✓ Can detect magnitude of changes if used without scaling
× Assumes linearity	× Less powerful	× Not scale invariant: results depend on scaling used
× Susceptible to outliers	× Ignores pattern of up- or down-regulation in time series	× Cannot detect negative correlations

8.3 Riduzione della Dimensionalità

Spesso, noi desideriamo visualizzare i dati del microarray, sia come aiuto a una analisi visuale, oppure come un'analisi preliminare rispetto all'applicazione di algoritmi più sofisticati. Il cervello umano si è evoluto in modo tale da essere capace di visualizzare oggetti a due o tre dimensioni: noi viviamo in un mondo a tre dimensioni, e vediamo per mezzo di una sorta di combinazione stereoscopica di immagini a due dimensioni che provengono dai nostri occhi. I nostri strumenti principali per visualizzare i dati di un microarray sono essi stessi bi-dimensionali: schermi dei computer, proiettori di immagini, articoli di ricerca e libri. Tutto ciò rende difficoltosa la visualizzazione dei dati dei microarrays: noi stiamo tentando di rappresentare dati altamente dimensionali in due o tre dimensioni .

Questo paragrafo descrive due metodi per visualizzare i dati dei microarray in due o tre dimensioni: *Analisi delle Componenti Principali e Scaling Multidimensionale.*

Analisi delle Componenti Principali

Immaginiamo una nuvola di punti nello spazio a tre dimensioni. Ora immaginiamo di porre una cartolina dietro i punti e guardiamo all'ombra dei punti sulla cartolina: noi abbiamo proiettato un gruppo di punti a tre dimensioni su uno spazio bidimensionale.

L'Analisi della Componente Principale (PCA) è un metodo che proietta uno spazio altamente dimensionale in uno spazio ad un numero minore di dimensioni. Noi abbiamo scelto un angolo con cui guardare allo spazio altamente dimensionale in modo tale da catturare molta della variabilità dei dati originali come se fossimo in uno spazio a minori dimensioni ed ignorare le altre dimensioni.

Esempio 8.7: analisi delle componenti principali di una penna

Supponiamo di avere una penna (ovviamente a tre dimensioni), e che si voglia costruire una visione bidimensionale di questa penna. Se noi guardiamo alla penna da una delle parti terminali, noi vedremmo in due dimensioni un cerchio con una protuberanza (Figura 8.8a).

Se noi ruotiamo la penna in modo tale da guardare lungo la sua lunghezza, ma facciamo in modo da nascondere il clip (gancetto), possiamo vedere che abbiamo un oggetto lungo, che è più grande sul lato del cappuccio (Figura 8.8b). In questa rotazione

abbiamo risolto la prima componente principale cercando l'asse con la massima variabilità nella forma della penna: il lato più lungo.

Se noi ruotiamo ancora la penna, in modo tale che si possa vedere il gancetto, esso è ora riconoscibile come una penna (Figura 8.8c). In questa seconda rotazione, abbiamo risolto la seconda componente principale: abbiamo trovato l'asse che è perpendicolare al primo asse che contiene la maggior parte della variabilità rimanente.

Figura 8.8: Analisi delle componenti principali di una penna. (a) Penna guardata da un'estremità. Nessun componente principale è stato identificato. **(b)** Penna guardata lateralmente con il gancetto nascosto. La prima componente principale è stata identificata; noi possiamo vedere che l'oggetto è lungo ed è più largo da un lato, ma il gancetto è oscurato. **(c)** Penna guardata lateralmente sul lato da cui si vede il gancetto; essa dovrebbe essere riconosciuta come penna. La seconda componente principale è risolta. Possiamo ora vedere che il gancetto e l'oggetto è riconoscibile come penna. Anche se in (c) qualche informazione è perduta. Noi non sappiamo che la penna è rotonda; potrebbe trattarsi di una penna quadrata.

Come funziona il PCA

Supponiamo che si voglia ridurre un esperimento con microarray con 10.000 geni in due o tre dimensioni. Per prima cosa costruiamo un qualcosa che è conosciuto come la matrice della *varianza-covarianza* per questi geni. Questa matrice cattura la variabilità di ciascun gene, e lo spazio n cui essa co-varia (equivalente alla correlazione) con ogni altro gene. Così noi avremmo un array con 10.000 x 10.000 elementi. Utilizziamo questo array per identificare una nuova variabile che è la combinazione lineare dei geni e che presenta la massima quantità di varianza. Questo è il primo componente principale (Figura 8.9).

Noi quindi troviamo la variabile che è ortogonale a questa prima variabile e che massimizza la rimanente varianza. Questo è il secondo componente principale. Ripetiamo il processo fino a quando abbiamo così tanti componenti principali quanti ce ne interessano.

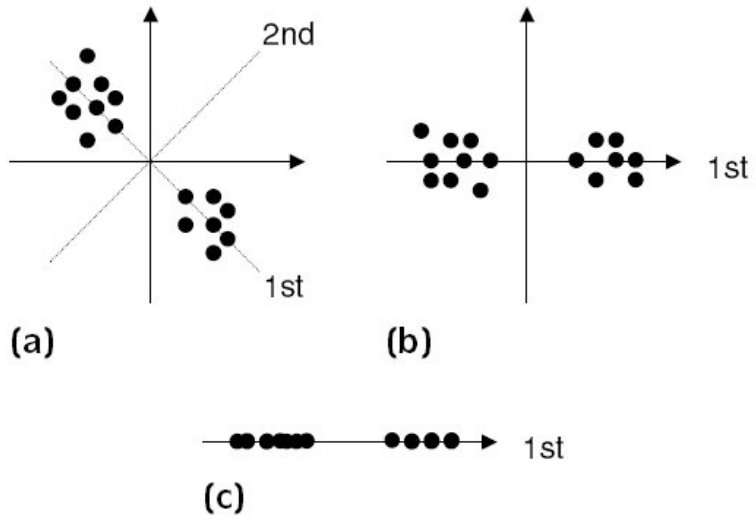


Figura 8.9: Analisi delle componenti principali su un semplice, immaginario data set a due dimensioni. Il data set è a due dimensioni e consiste in due cluster ellittici. **(a)** le componenti non sono risolte. **(b)** la prima componente è lungo la direzione della massima variabilità, in cui il cluster dei punti mostra una separazione. Poiché il cluster è soltanto a due dimensioni, noi non abbiamo scelta per quanto riguarda la seconda componente, che deve essere ortogonale alla prima. **(c)** Noi possiamo risolvere questo insieme di dati bidimensionali in un solo insieme di dati ad una dimensione e vedere il cluster dei dati.

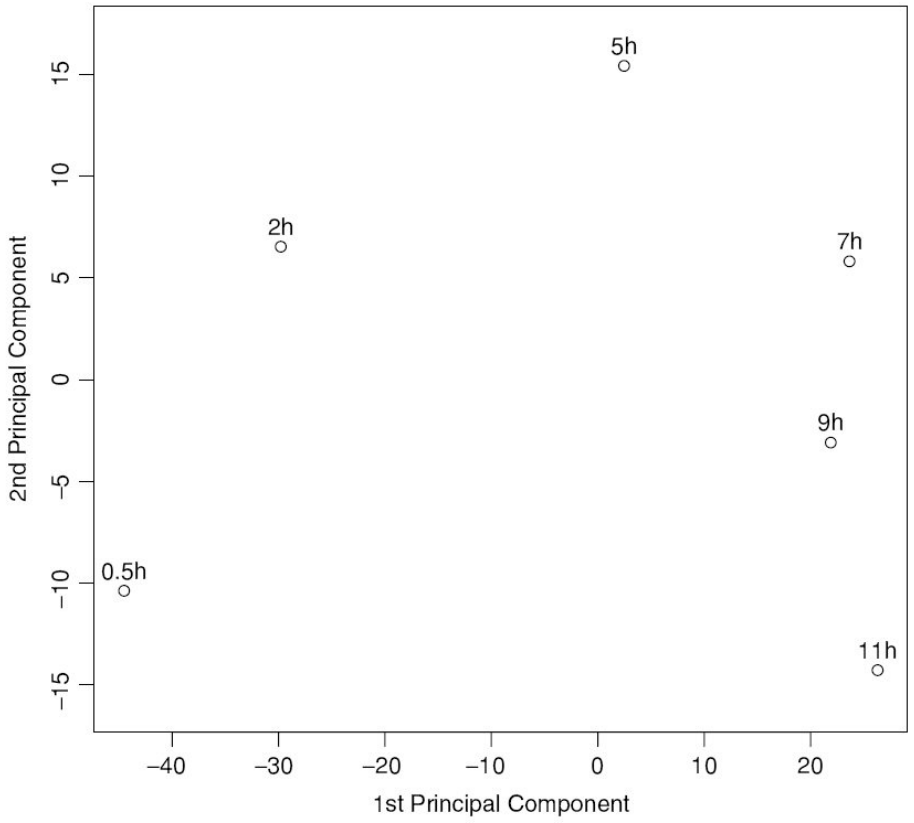


Figura 8.10: Analisi delle componenti principali dei dati di sporulazione del lievito. L'analisi delle componenti principali sui 1000 geni maggiormente variabili dal data set 8A nel tentativo di essere in grado di visualizzare la similarità nell'ambito di sei esempi. Il diagramma mostra un certo numero di interessanti caratteristiche:

- I sei campioni mostrano un chiaro andamento nel diagramma delle componenti principali; l'analisi ha un senso biologico.
- Se noi guardiamo solo alla prima componente principale, i campioni dopo 7, 9 ed 11 ore appaiono tutti raggruppati insieme, mentre i primi campioni si risolvono (in ordine). Questo implica che i tre ultimi campioni sono del tutto simili l'uno all'altro, e che i cambiamenti più importanti della trascrizione avvengono entro le prime 7 ore.
- La seconda componente principale risolve i tre ultimi punti temporali; c'e un fatto molto interessante: gli ultimi punti temporali hanno valori simili ai primi. Questo fatto suggerisce che vi possano essere processi transitori che si accendono nel'arco delle 5 ore, e che poi si spengono.

Esempio 8.8: analisi del componente principale dei dati della sporulazione del lievito

Vogliamo vedere la relazione che intercorre tra sei punti temporali del data set 8A. Non possiamo visualizzare 6000 geni, quindi usiamo la PCA per visualizzare la relazione tra i punti temporali in un diagramma a due dimensioni. Questo è un buon esempio illustrativo poiché ci aspettiamo a priori che punti temporali vicini dovrebbero essere simili.

In questo esempio, il PCA rivela tre interessanti aspetti relativi ai dati (Figura 8.10);

- Punti temporali vicini sono vicini l'uno all'altro nella figura; possiamo perfino vedere "l'andamento" intrapreso dalle serie temporali attraverso la spazio delle componenti principali. Da questo fatto concludiamo che l'analisi abbia senso biologicamente.
- Se noi guardiamo alla prima componente principale (asse x), i campioni dopo 7, 9 e 11 ore, sono raggruppati insieme, mentre gli altri tre punti temporali sono risolti, in ordine, lungo l'asse.
- La seconda componente principale risolve i tre successivi punti temporali ma - fatto interessante - la tendenza nella seconda componente principale è una salita in valore dei campioni durante le 5 ore, e quindi una ridiscesa, in modo tale che gli ultimi campioni siano molto simili ai primi.

Il comportamento di queste due componenti suggerisce due serie di processi: una serie di processi che cambia durante il periodo di 7 ore e persiste nello stato alterato durante la sporulazione (corrispondente al primo componente principale), ed un altra serie di processi che sono attivati in un periodo di 5 ore, ma che non ritornano al loro stato iniziale (corrispondente alla seconda componente principale). In questo modo, semplicemente guardando al diagramma del componente principale, noi possiamo acquisire una comprensione dei fenomeni biologici che sono alla base di quanto sopra.

La PCA inoltre, identifica la quantità di variabilità catturata in ciascuno dei componenti (Tabella 8.3). In questo esempio, la maggior parte della variabilità (77%) è catturata dalla prima componente, e l'88% dalle successive due; 5 componenti catturano la totalità della variabilità dei dati. Noi, pertanto, ci aspettiamo che questo esempio sia relativamente semplice dal punto di vista biologico, con un piccolo numero di modi di evoluzione attivi durante la sporulazione. Con dati più complessi, la variabilità può essere distribuita su parecchi componenti principali.

TABLE 8.3: Principal Components of Data Set 8A

Principal Component	1st	2nd	3rd	4th	5th
Standard deviation	30.3	11.3	9.1	5.9	5.3
Proportion of variance	77%	11%	7%	3%	2%
Cumulative proportion	77%	88%	95%	98%	100%

Nota: I primi cinque componenti principali spiegano essenzialmente tutto sulla variabilità nei 1.000 geni usati per questa analisi. Così benché noi si guardi ad un numero di geni molto grande, non vi sono molti differenti processi in questo esperimento: l'88% della variabilitá è spiegata dai due componenti usati per la Fig. 8.10. Pertanto, questa figura è una buona rappresentazione delle similarità e delle differenze circa i sei campioni.

Scaling Multidimesionale

Lo Scaling Multidimensionale (MDS) è un approccio differente alla riduzione della multidimensionalità e alla visualizzazione. A differenza della PCA, esso non parte dai dati, ma piuttosto dalle misure di distanza tra i campioni, oppure tra i profili che stiamo confrontando. Noi misuriamo la distanza tra i profili usando una qualsiasi delle misure descritte ne Paragrafo 8.2. MDS, quindi, tenta di localizzare i profili nello spazio a due - o tre dimensioni - in modo tale che le distanze nello spazio, bidimensionale o tridimensionale, siano il più possibile vicine alle distanze misurate tra i profili nello spazio a dimensioni più elevate.

Esempio 8.9: Scaling Multidimensionale sui dati di sporulazione del lievito

Noi sviluppiamo l'MDS sui dati del data set 8A; in questo modo confrontiamo i risultati dell'MDS con i risultati del PCA sui stessi dati. Uno dei vantaggi del MDS rispetto al PCA è che noi misuriamo la distanza tra i campioni in differenti modi: In questo esempio, abbiamo calcolato la matrice delle distanze usando sia la correlazione della distanza di Pearson, che la distanza Euclidea (Tabella 8.4).

TABLE 8.4: Multidimensional Scaling of Data Set 8A

	Measured Distance Between Profiles (Scaled Distance in Parentheses)				
	0.5 h	**2 h**	**5 h**	**7 h**	**9 h**
Euclidean distance					
2 h	31 (22)				
5 h	55 (54)	38 (33)			
7 h	71 (70)	58 (53)	31 (23)		
9 h	68 (67)	55 (53)	31 (27)	22 (9)	
11 h	73 (71)	61 (60)	39 (38)	30 (20)	21 (12)
Correlation distance					
2 h	0.23 (0.26)				
5 h	0.53 (0.53)	0.28 (0.33)			
7 h	0.53 (0.54)	0.36 (0.38)	0.07 (0.12)		
9 h	0.63 (0.63)	0.45 (0.46)	0.12 (0.15)	0.04 (0.10)	
11 h	0.67 (0.68)	0.53 (0.53)	0.18 (0.24)	0.09 (0.15)	0.06 (0.09)

Nota: L'MDS dei sei profili inizia con la matrice delle distanze, con le distanze misurate tra i profili e trova i punti nello spazio in uno spazio a due (o tre) dimensioni in modo tale che le distanze tra quei punti siano più vicini alle distanze nella matrice. Quei punti sono indicati in Figura 8.9. Le distanze attuali sono mostrate con le distanze tra i punti nello spazio bidimensionale mostrato tra parentesi. Alcune di queste sono molto vicine, per esempio, le distanze tra i campioni dopo 0.5 ore. Quando le distanze sono piuttosto differenti (e.g., la distanza tra campioni dopo 9 ed 11 ore), si evince che il mapping dei dati in due dimensioni non è accurato. Alcune informazioni sono state perdute e richiederebbero una extra dimensione per essere visualizzate. Un modo di pensare a questo fatto è che sia i campioni a 9 ore che i campioni ad 11 ore vogliono uscire fuori dalla pagina ed essere in qualche parte nello spazio al di sopra del libro.

Usando queste misure di distanza, noi troviamo i punti nello spazio bidimensionale che hanno distanza simile fra di loro (Figura 8.11). Il diagramma MDS usando la distanza Euclidea è quasi identico al diagramma PCA. Il diagramma MDS che fa uso della correlazione presenta alcune differenze: i primi punti temporali sono più sparsi, mentre gli ultimi punti sono raggruppati in modo più vicino l'uno all'altro. In entrambi i casi, la struttura inerente le serie temporali è preservata, ed i suoi punti sono in accordo ad un "andamento" ben riconoscibile.

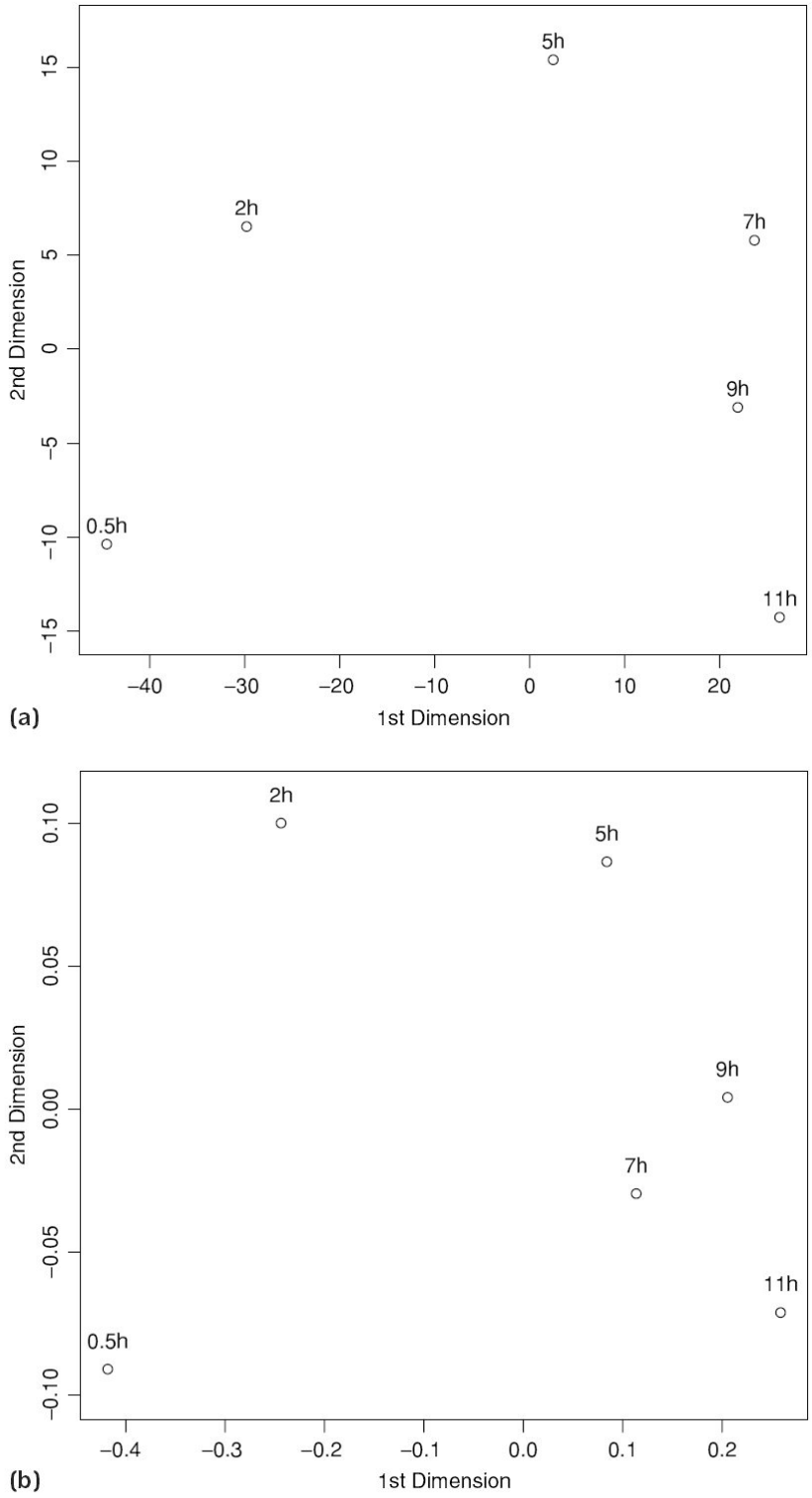


Figura 8.11: Scaling Multidimensionale dei dati della sporulazione del lievito. In entrambe le figure, noi abbiamo mappato i sei punti temporali dentro uno spazio a due dimensioni così che la distanza tra i punti a due dimensioni sono il più vicino possibile alle distanze tra i profili usando 1000 geni fortemente variabili.

(a) Viene usata la distanza Euclidea. Il diagramma è quasi identico al diagramma PCA e mostra la stessa chiara progressione lungo i punti in direzione del tempo. **(b)** Viene usata la Correlazione come misura della distanza. Il diagramma è simile, ma c'è qualche differenza: la distanza tra i primi punti temporali è elevata, mentre gli ultimi tre campioni sono raggruppati molto vicini l'un l'altro. L'Andamento naturale nello spazio bidimensionale visto con il PCA e con la distanza Euclidea MDS non è presente per i campioni dopo 7, 9 ed 11 ore.

L'MDS è un eccellente metodo molto naturale per visualizzare la matrice delle distanze tra i geni o profili di geni. Gli algoritmi di clustering descritti nei rimanenti paragrafi di questo capitolo lavorano tutti con la matrice delle distanze così che l'analisi della struttura del cluster che si sviluppa dovrebbe riflettere il diagramma MDS. Questo ci può aiutare a determinare la misura della distanza, che poi viene usata con l'analisi del cluster. MDS può anche essere usata come guida per la scelta del numero di cluster per i K-means clustering che descriveremo nel Paragrafo 8.6.

Il diagramma MDS ha il vantaggio rispetto al diagramma dei cluster di non imporre nessuna struttura dei dati. Esso ha lo svantaggio di essere difficile da usare per un elevato numero di geni. Riassumiamo i vantaggi e gli svantaggi del PCA rispetto all'MDS in Tabella 8.5.

TABLE 8.5: Comparison of Principal Component Analysis with Multidimensional Scaling

Principal Component Analysis	Multidimensional Scaling
✓ Can visualise high-dimensional data in two or three dimensions	✓ Can visualise high-dimensional data in two or three dimensions
✓ Do not impose any a priori structure on the relationships between genes and samples	✓ Do not impose any a priori structure on the relationships between genes and samples
✓ Can be used as inputs for classification techniques in Chapter 9	✓ Can resolve non-linear relationships if used with a non-linear distance measure
✓ Implemented in wide range of packages, including GeneSpring, J-Express, R and Matlab	✓ Allows visualisation of distance matrix to be used for cluster analysis and can be used to help select an appropriate distance measure
× Principal components are abstract concepts and have no concrete meaning	✓ Can be used as inputs for classification techniques in Chapter 9
× Can only resolve linear relationships between genes and samples	✓ Implemented in R and Matlab
× Susceptible to outliers – uses raw, scaled or centred data	× Dimensions have no meaning at all
× Difficult to visualise large numbers of genes or samples	× Different distance measures give different results
	× Difficult to visualise large numbers of genes or samples
	× Not currently implemented in GeneSpring, J-Express or other commonly used gene expression analysis packages

8.4 Clustering Gerarchico

I prossimi due paragrafi discuteranno il tool di analisi per l'espressione del gene più largamente usato: il clustering gerarchico. Questa è una metodologia che organizza i geni o i profili campione secondo una struttura ad albero in modo tale che i profili simili appaiano vicini l'un l'altro, mentre i profili dissimili appaiano piuttosto distanti l'un l'altro.

La tecnica è diventata popolare per le seguenti ragioni:

- Essa può semplificare grandi volumi di dati.

- L'Analisi rivela gruppi di geni similari che possono essere studiati in grande dettaglio.
- È possibili visualizzare i dati in modo gerarchico utilizzando programmi di computer interattivi.
- I risultati sono visualizzati nello stile delle analisi filogenetiche, che sono familiari a molti genetisti.

I questo paragrafo, noi descriveremo come il clustering gerarchico lavori, applicandolo prima all'analisi di un piccolo numero di geni, e quindi ad un numero di geni sempre più largo. Poniamo, dunque, la nostra attenzione ad un certo numero di considerazioni metodologiche del clustering gerarchico ed agli effetti derivanti dall'uso di differenti misure di distanza. Il Paragrafo 8.5 descrive un metodo per determinare l'affidabilità e la robustezza dei cluster in relazione alla variabilità ed al rumore negli esperimenti con microarray.

Esempio 8.10: il metodo base per il Clustering Gerarchico

Mostreremo il metodo di clustering gerarchico per mezzo di un semplice esempio. Raggrupperemo 5 geni dal data set 8B: CREME9, ALOX5, HS2ST1, PELI1 e RDHL.

In Tabella 8.6, mostriamo la matrice delle distanze per questi 5 geni. Questa è stata calcolata mediante la correlazione di Spearman. L'Algoritmo ha 4 passi:

1. Esaminare la matrice della distanza e trovare le entrate più vicine (che possono essere geni o cluster di geni).
2. Congiungere queste entrate insieme nell'albero e formare un nuovo cluster.
3. Calcolare la distanza tra il cluster formato or ora, e gli altri geni e il cluster.
4. Ritornare al punto 1 e ripetere il procedimento fino a quando tutti i geni ed i cluster sono intercollegati.

Se noi guardiamo alla tabella 8.6a, noi possiamo vedere che i due geni più vicini sono CREME9 e RDHL. Pertanto, questi sono i primi due geni ad essere intercollegati (Figura 8.12a), formando un nuovo cluster che contiene questi due geni. Procediamo verso il passo 3, nel quale calcoliamo la distanza tra i rimanenti geni ed il cluster contenente CREME9 e RDHL (Tavola 8.6). Vi è un certo numero di metodi differenti per fare ciò, e l'albero che si viene formando dall'uso di questi metodi appare di frequente differente. Qui, la distanza è stata calcolata usando un **linkage medio**: noi discutiamo questo ed altri metodi nel seguito. Ritorniamo ora sul passo 1 dell'algoritmo: le entrate più vicine nella tabella sono i geni ALOX5 e PELI1. Questi vengono intercollegati per formare un nuovo cluster (Figura 8.12b), e noi ritorniamo ancora al passo 1. Il processo continua fino a quando tutti i geni ed i cluster sono combinati insieme (Tabella 8.6; Figura 8.12). Il risultato è esattamente il tipo di albero che si potrebbe vedere nei programmi, quali TreeView, GeneSpring e J-Express. Tradizionalmente, l'altezza dei tre rami è proporzionale alla distanza tra i geni o i cluster. Pertanto, i geni più vicini (e.g., CREME9 e RDHL) saranno congiunti da rami più corti che non i geni distanti (e.g., ALOX5 e PELI1), che sono connessi da rami più lunghi.

TABLE 8.6A: Hierarchical Clustering

Gene	CREME9	ALOX5	HS2ST1	PELII	RDHL
CREME9	0.00				
ALOX5	0.57	0.00			
HS2ST1	0.79	0.46	0.00		
PELI1	0.39	0.39	0.51	0.00	
RDHL	0.03	0.62	0.79	0.43	0.00

Nota: Le distanze tra ciascuno dei cinque geni dal data set 8B sono calcolate. Nel primo passo del processo, noi identifichiamo la distanza più piccola: tra CREME9 e RDHL. Questa è una matrice triangolare; la distanza tra i due geni è la stessa da qualunque punto di vista si guardi ad essi.

TABLE 8.6B: Hierarchical Clustering

Gene	CREME9-RDHL	ALOX5	HS2ST1	PELII
CREME9-RDHL	0.00			
ALOX5	0.59	0.00		
HS2ST1	0.79	0.46	0.00	
PELI1	0.41	0.39	0.51	0.00

Nota: I geni CREME9 e RDHL sono stati combinati per formare un singolo cluster. Noi calcoliamo la distanza tra questo cluster e ciascuno degli altri geni. Vi è un certo numero di metodi per questo genere di calcoli che sono discussi nel capitolo 8.4; qui noi usiamo un metodo denominato linkage medio. Le entrate più vicine nella tavola sono ora tra ALOX5 e PELI1.

TABLE 8.6C: Hierarchical Clustering

Gene	CREME9-RDHL	ALOX5-PELI1	HS2ST1
CREME9-RDHL	0.00		
ALOX5-PELI1	0.50	0.00	
HS2ST1	0.79	0.46	0.00

Nota: È' stato ora formato un nuovo cluster contenente i geni ALOX5 e PELI1. Noi calcoliamo la distanza tra questi cluster ed i rimanenti geni. Le entrate più vicine sono ora tra il cluster contenente ALOX5 e PELI1, ed il gene H2ST1.

TABLE 8.6D: Hierarchical Clustering

Gene	CREME9-RDHL	ALOX5-PELI1-HS2ST1
CREME9-RDHL	0.00	
ALOX5-PELI1-HS2ST1	0.60	0.00

Nota: Nel passo finale, I due cluster rimanenti saranno uniti l'uno all'altro.

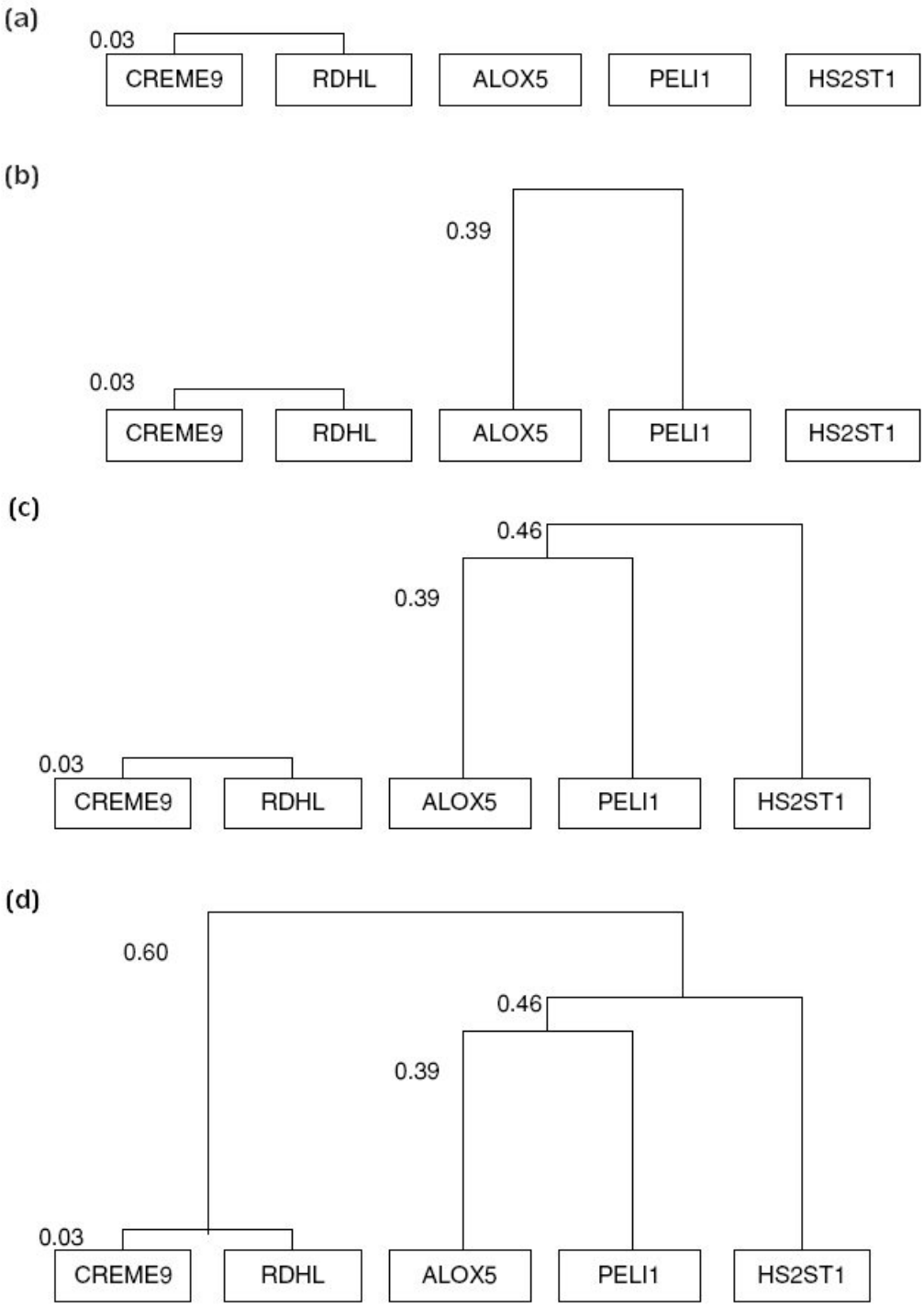


Figura 8.12 Clustering gerarchico. La Cluster analisi è applicata a 5 geni dal data set 8B; la matrice distanza per questi geni è mostrata nella Tabella 8.6. Vi sono 4 passi: **(a)** i due geni più vicini, CREME9 e RDHL, sono intercollegati per formare un cluster; la distanza tra di essi è 0.03. L'Algoritmo continua con il passo 2, in cui calcoliamo la distanza tra ciascuno dei rimanenti geni ed il cluster contenente CREME9 e RDHL. **(b)** La successiva distanza più piccola è tra ALOX5 e PELI1; la distanza è 0.39. Questi sono congiunti per formare un nuovo cluster. **(c)** la successiva più piccola distanza è tra il cluster contenente ALOX5 e PELI1, ed il gene HS2ST1, una distanza di 0.46. Questi sono congiunti per formare un cluster a tre geni contenente ALOX5, PELI1 e HS2ST1. Il cluster contenente ALOX5 e PELI1 forma un sottocluster del cluster a tre geni. **(d)** I due rimanenti cluster sono congiunti per completare l'albero. Abbiamo contrassegnato le distanze sull'albero; tradizionalmente, le altezze dei rami sono proporzionali alla distanza tra i geni, in modo tale che geni molto vicini dovrebbero avere collegamenti molto corti, mentre geni distanti dovrebbero avere lunghi collegamenti.

Esempio 8.11: Clustering Gerarchico su un data set più grande

Applichiamo il clustering gerarchico a 15 geni dal data set 8A. Questi geni hanno parecchie funzioni: riparazione del DNA, riparazione per escissione di nucleotidi, biosintesi delle proteine, risposta allo stress, inizio della trascrizione ed altre funzioni sconosciute.

Dai profili delle serie temporali (Figura 8.13), si può vedere che alcuni geni potrebbero essere raggruppati insieme, per esempio, CDC21 e DIN7 hanno forma molto simile, benché in scale lievemente diverse.

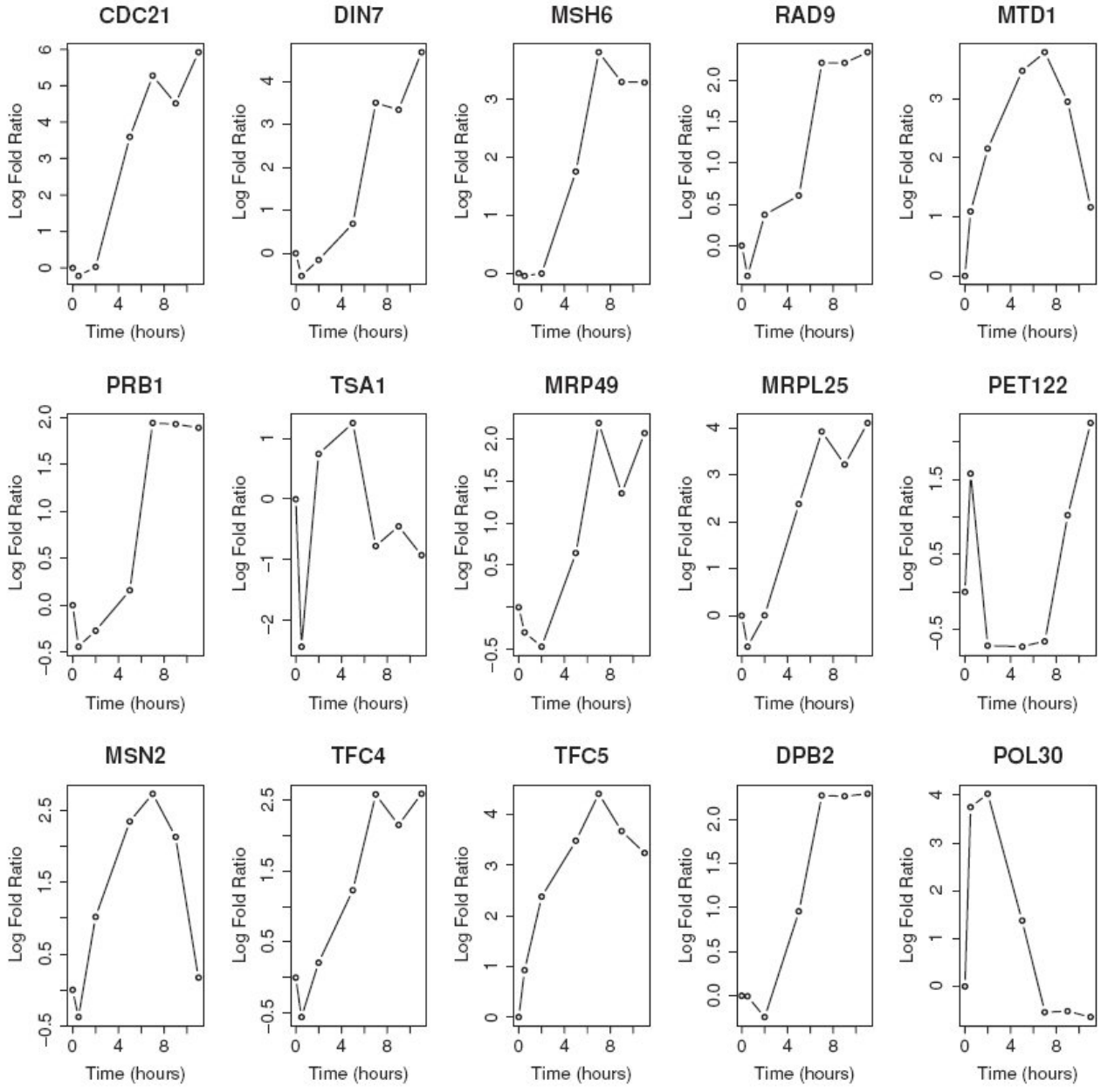
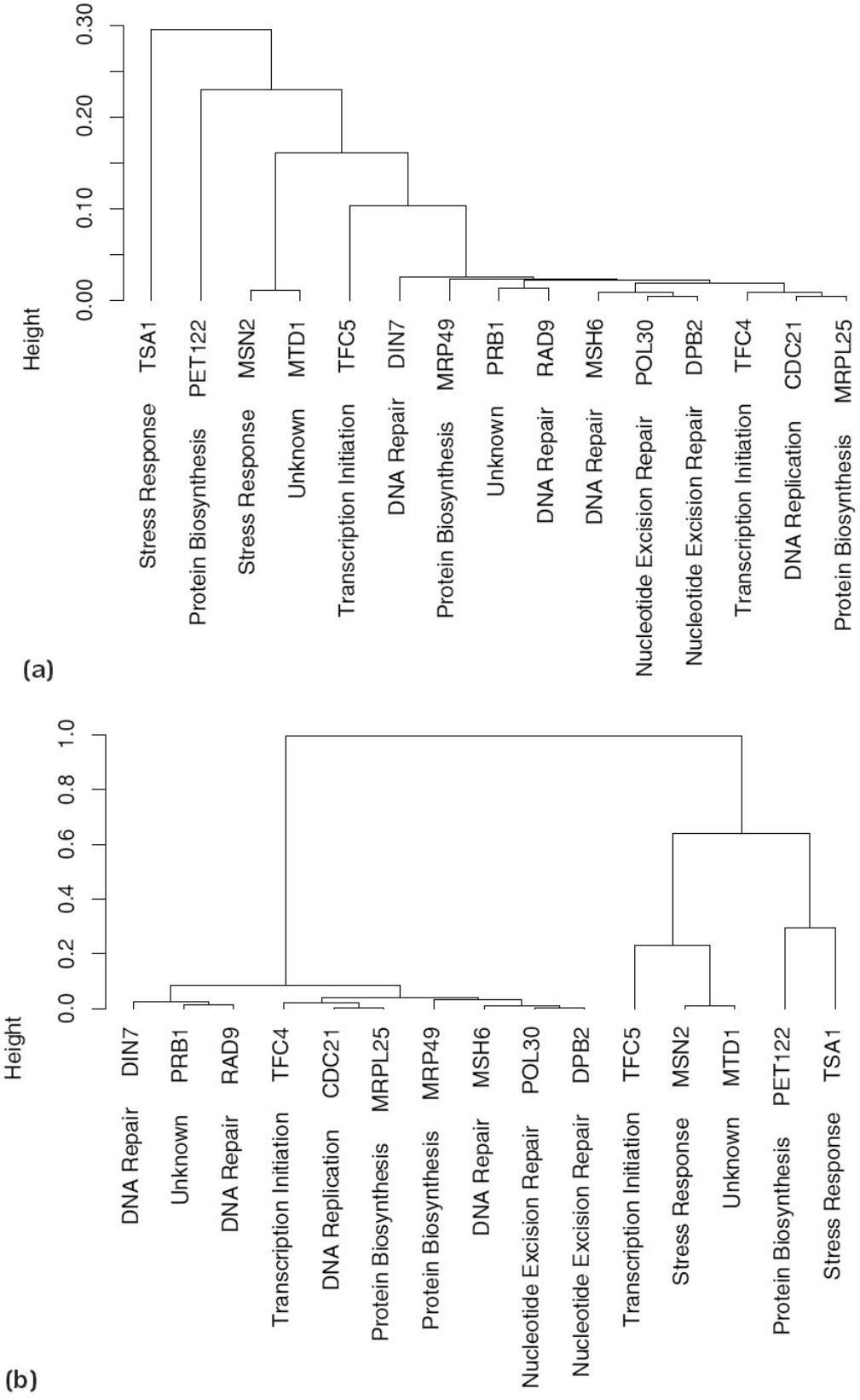
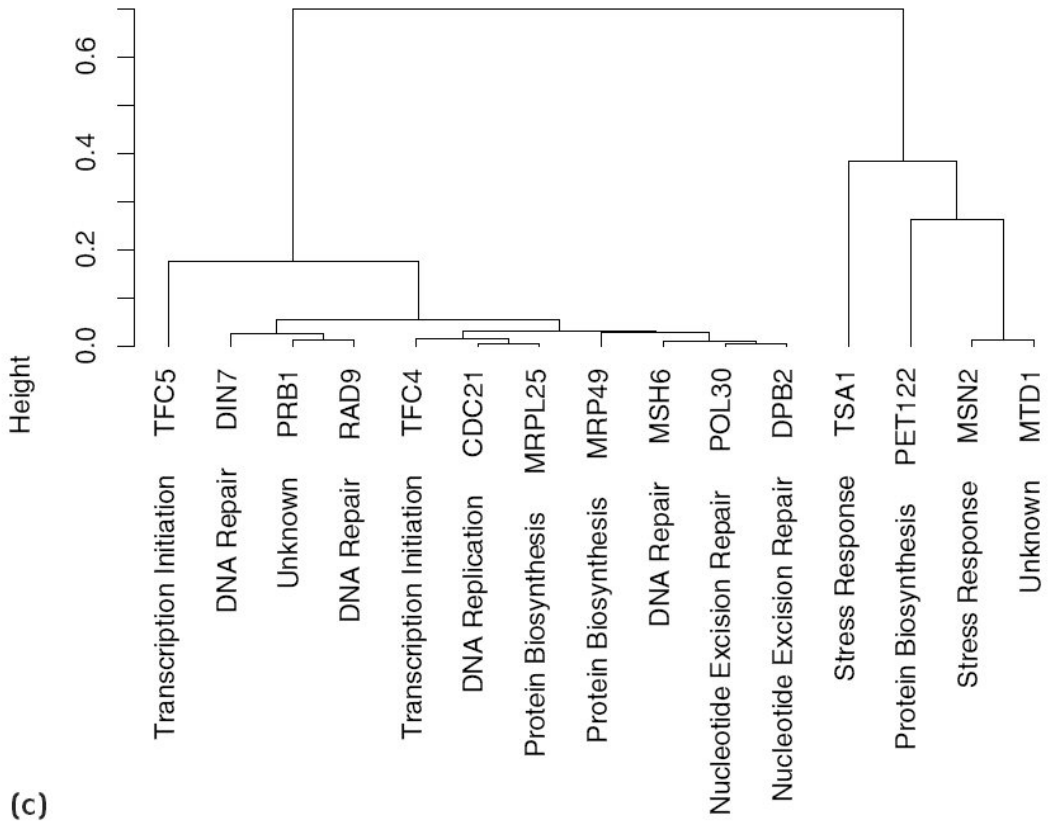


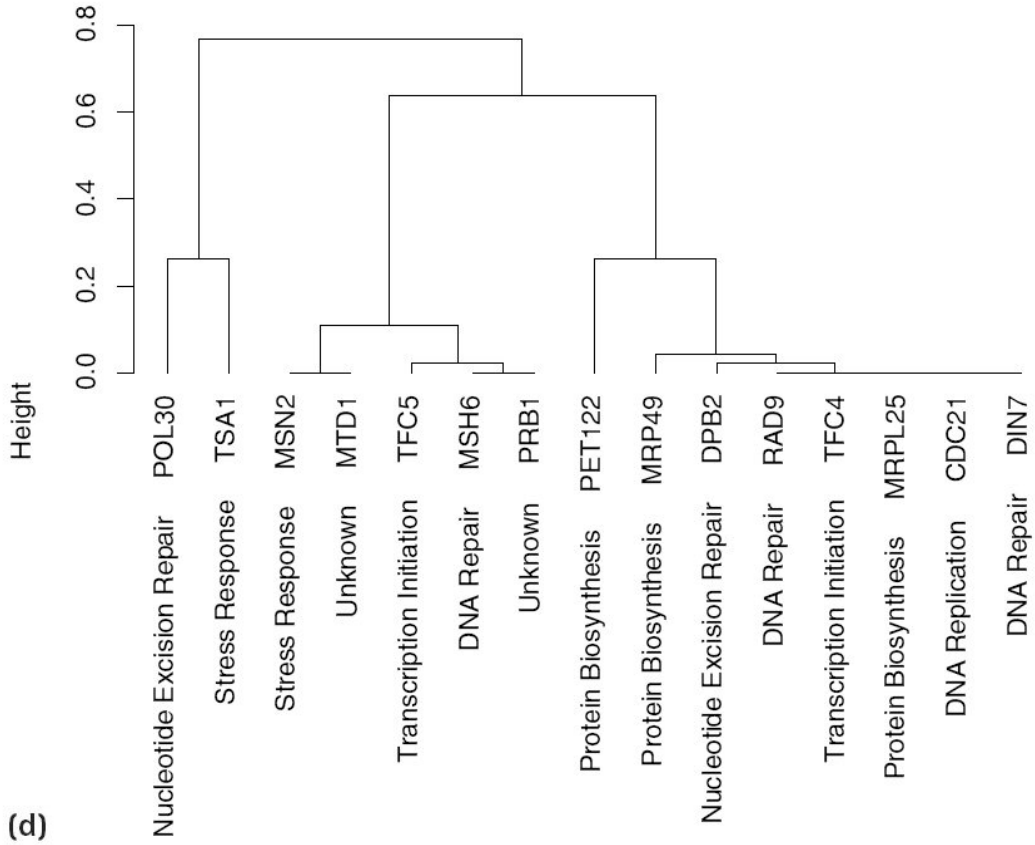
Figura 8.13: Profili di geni per il clustering. Mostriamo le serie temporali per 15 geni dal data set 8A di cui abbiamo fatto il cluster nel'esempio sulla analisi del clustering. Questi geni hanno un certo numero di funzioni differenti: Riparazione del DNA (DIN7, MSH6, RAD9), replicazione del DNA (CDC21), riparazione per escissione di nucleotidi (DPB2, POL30), biosintesi delle proteine (MRP49, MRPL25, PET122), inizio della trascrizione (TFC4, TFC5), risposta allo stress (TSA1, MSN2) ed alcune funzioni sconosciute (MTD1 e PRB1). Guardando ai profili, possiamo vedere alcuni cluster "naturali": geni la cui trascrizione è persistentemente down-regolata o up-regolata, e geni con transientemente up - o down - regolati. Tutte le misure di espressione del gene sono rapporti relativi all'espressione al tempo zero. Abbiamo incluso il punto (0,0) in tutti i diagrammi, ma non lo abbiamo incluso nell'analisi poiché esso non è un valore misurato.

La Figura 8.14c è un dendrogramma che è stato costruito usando la correlazione di Pearson e il linkage medio. Il cluster dei geni si suddivide in due grandi gruppi: quelli che sono persistentemente up-regolati nel corso del tempo, e quelli che mostrano una risposta transitoria. Si noti che la correlazione di Pearson rivela correlazioni negative, così che il PET1222 è ragionevolmente vicino a MSN2 ed a MTD1, mentre il gene POL30 è molto vicino ai geni DIN7 e MSH6.




Height
0.0
0.2
0.4
0.6
TFC5
DIN7
PRB1
RAD9
TFC4
CDC21
MRPL25
MRP49
MSH6
POL30
DPB2
TSA1
PET122
MSN2
MTD1
Transcription Initiation
DNA Repair
Unknown
DNA Repair
Transcription Initiation
DNA Replication
Protein Biosynthesis
Protein Biosynthesis
DNA Repair
Nucleotide Excision Repair
Nucleotide Excision Repair
Stress Response
Protein Biosynthesis
Stress Response
Unknown

(c)


Height
0.0
0.2
0.4
0.6
0.8
POL30
TSA1
MSN2
MTD1
TFC5
MSH6
PRB1
PET122
MRP49
DPB2
RAD9
TFC4
MRPL25
CDC21
DIN7
Nucleotide Excision Repair
Stress Response
Stress Response
Unknown
Transcription Initiation
DNA Repair
Unknown
Protein Biosynthesis
Protein Biosynthesis
Nucleotide Excision Repair
DNA Repair
Transcription Initiation
Protein Biosynthesis
DNA Replication
DNA Repair

(d)

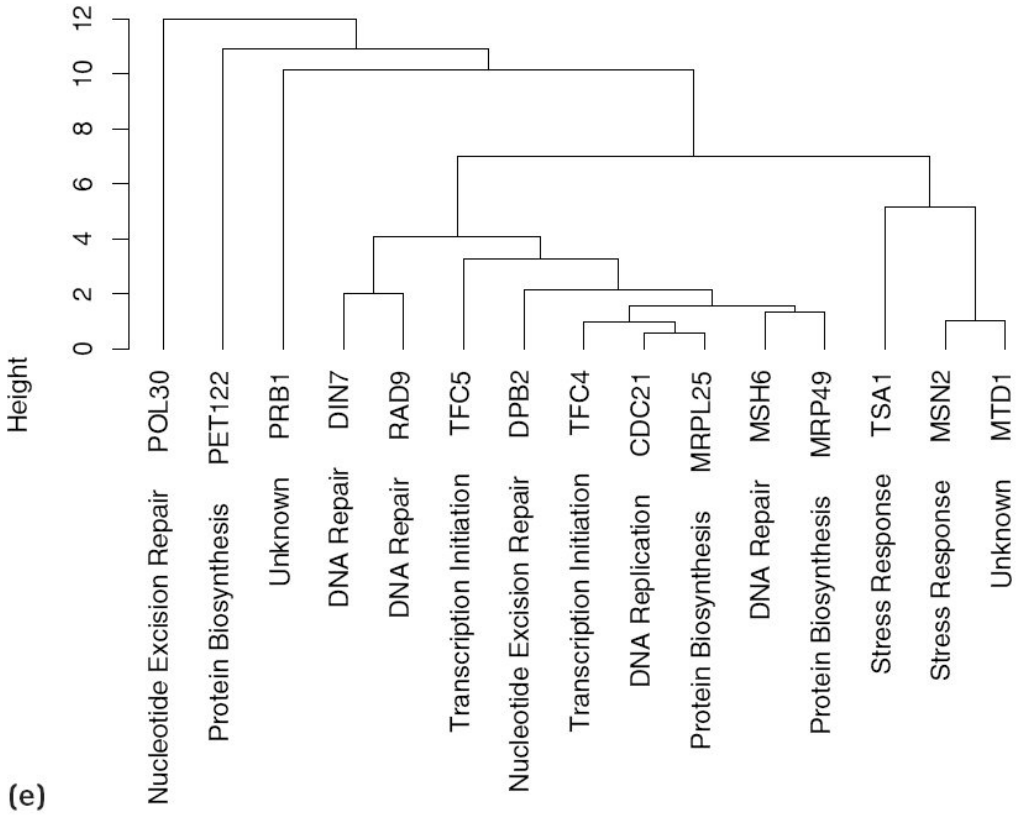


Figura 8.14: Clustering Gerarchico. (a) Clustering con la correlazione di Pearson e linkage singolo. Il linkage singolo tende a produrre *catene*, che vediamo nella figura. L'Algoritmo ha trovato il grande cluster di geni persistenti, ma i geni transitori sono addizionati al grande cluster uno alla volta, piuttosto che formare un proprio cluster. Il linkage singolo non è di solito un buon metodo per l'analisi dei microarray. **(b)** Cluster con la correlazione di Pearson e linkage completo. I due cluster principali sono ben definiti. La differenza più ovvia tra questo sistema di clustering ed il linkage medio è l'inclusione di TFC5 con i geni transitori.
(c) Clustering con la correlazione di Pearson e il linkage medio. Questo è un metodo usato molto comunemente per i dati dei microarray. Vi sono due grandi cluster corrispondenti alle risposte persistenti ed alle risposte transitorie. PET122 viene a far parte del cluster dei transitori. TFC5 viene a far parte del cluster dei geni persistenti, ma giace al di fuori del gruppo principale. Si noti che TFC5 potrebbe esser stato disegnato tra i due grandi cluster, che probabilmente sarebbe il posto più giusto in cui allocarlo. Il cluster grande presenta un notevole grado di struttura fine. **(d)** Clustering con la correlazione di Spearman e linkage medio. Parecchi geni hanno distanza zero tra di essi, e quindi non esiste struttura fine. I gruppi sono abbastanza differenti dai gruppi con la correlazione di Pearson. MSH6 e PRB1 sono più vicini ai geni transitori poiché l'espressione del gene decresce molto lentamente. **(e)** Clustering con la distanza Euclidea e il linkage medio. Questo è molto differente dagli altri clustering. Un fatto molto importante è che i due geni sono negativamente correlati con gli altri profili, PET122 e POL30 sono fuori range e non sono raggruppati con nessun altro gene. I gruppi transitori e persistenti sono identificati, e vi è una distanza più grande tra i geni e meno struttura fine.

Metodi di collegamento (Linkage)

Nel passo 3 dell'algoritmo, congiungiamo due geni (oppure cluster) insieme, per formare un nuovo cluster, ed abbiamo bisogno di calcolare la distanza tra i nuovi cluster ed i rimanenti geni (o cluster). Esistono un certo numero di metodi che possono essere usati per calcolare le nuove distanze. Ciascun metodo produrrà un differente clustering; quindi è di estrema importanza scegliere con cura il metodo da usare.

In questo paragrafo discuteremo i tre metodi usati più comunemente: **linkage singolo, linkage completo, linkage medio**. Vi sono molti altri metodi, e questi sono implementati in pacchetti statistici come R e SPSS, oltre che nei packges dedicati all'analisi dei microarrays, come GeneSpring e J-Express.

Il linkage singolo definisce la distanza tra due cluster come la distanza tra i punti più vicini tra i cluster (Figura 8.15a). Il clustering che usa il linkage singolo (Figura 8.14a) tende a produrre un effetto chiamato *chaining*: singoli geni sono addizionati al cluster uno alla volta. Questo puó essere visto nella porzione di sinistra di Figura 8.14a, dove i geni transitori sono addizionati al cluster principale uno alla volta. Il linkage singolo può essere utile quando i dati hanno già una sorta di cluster naturale, sono ben definiti ma hanno forma irregolare; ma esso non è generalmente raccomandato per i dati dei microarray.

Il linkage completo definisce la distanza tra due cluster come la distanza tra i punti più lontani tra i cluster (Figura 8.14b). Questo metodo funziona bene quando vi siano cluster ben definiti, mentre funzionano meno bene quando i dati siano sfumati (fuzzy). Il linkage medio definisce la distanza tra due cluster come media delle distanze tra tutte le coppie di punti tra i due cluster (Figura 8.15c). Il linkage medio sta in posizione intermedia tra il linkage singolo ed il linkage completo e si comporta molto bene in molte applicazioni con i microarray (Figura 8.14c).

L'importante proprietà di cui prendere atto circa questi ed altri metodi di linkage è che essi producono differenti diagrammi di cluster (Figura 8.14). Geni individuali possono essere riuniti in cluster in modo differente a seconda del metodo usato; per esempio, il gene TFC5 é raggruppato in modo differente in tre dendrogrammi. Pertanto, non è saggio inferire troppo da un dato cluster in ciascun dendrogramma poiché questi cluster possono essere presenti soltanto come risultato della metodologia scelta.

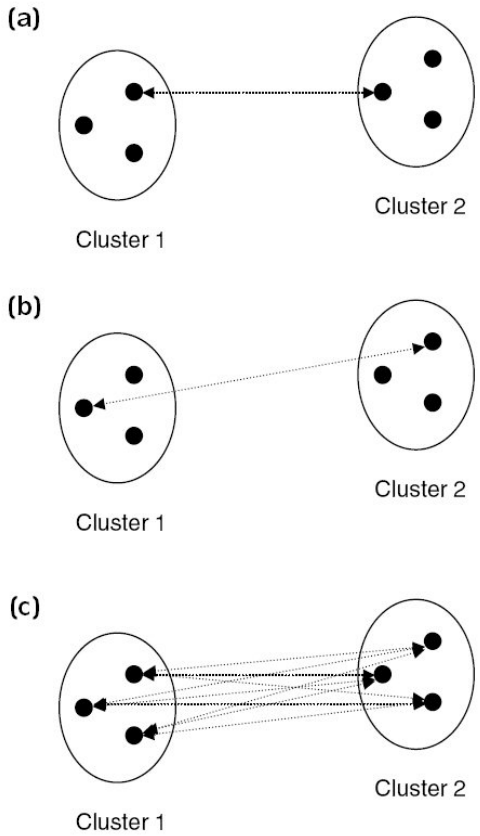


Figura 8.15: Metodi di linkage. (a) Linkage singolo. La distanza tra due cluster è definita come la distanza tra i due punti più vicini tra i due cluster. **(b)** Linkage completo. La distanza tra i cluster è definita come la distanza tra i punti più lontani tra i due cluster. **(c)** Linkage medio. La distanza tra due cluster è definita come la media delle distanze tra tutti i punti dei cluster.

Misure della Distanza

Noi abbiamo già visto che il metodo di linkage può produrre differenti cluster. La distanza metrica che si usa può anche risultare in differenti cluster. Dimostriamo ciò mostrando come il cluster sia diverso usando la correlazione di Pearson, la correlazione di Spearman e la distanza Euclidea. (Paragrafo 8.2).

La Figura 8.4 mostra i dendrogrammi dei 15 geni dell'esempio 8.11 che sono stati raggruppati utilizzando il linkage medio, ma con le tre misure di distanza descritte. Vi sono tre importanti caratteristiche da notare:

- **Correlazione negativa**. Le correlazioni di Pearson e Spearman hanno la capacità di individuare correlazioni negative, cioè geni con profili opposti. Un esempio è il gene PET122. Questo ha un profilo che è simile nella forma a MTD1 e a MSN2, ma che è in opposizione di fase. Quando PET122 è up-regolato, MTD1 e MSN2 sono down-regolati, e viceversa. Nei clustering prodotti dalla correlazione di Pearson e di Spearman, questi geni sono vicini. Ma con la distanza Euclidea, i profili sono molto distanti, sí da apparire molto separati nel dendrogramma.
- **I Cluster con più di un gene.** La correlazione di Spearman può produrre una distanza pari a zero se i profili dei geni hanno esattamente la stessa forma; la struttura fine del cluster scompare ed è sostituita con il cluster contenente un gruppo di geni. Un esempio è il cluster contenente MRP49, RAD9, TCF4, MRPL25, CDC21 e DIN7. Tutti questi hanno esattamente la stessa forma, e quindi hanno una distanza zero se viene applicata la correlazione di Spearman.
- **Distanze più grandi**. La distanza Euclidea tende a produrre distanze più grandi rispetto alla correlazione, così che i cluster sono "meno stringenti"

Quale misura di distanza conviene usare? Non esiste una risposta corretta. Noi raccomandiamo che si applichino tutte le misure di distanza nella nostra analisi dei cluster e guardare ai risultati ottenuti da tutti i metodi prima di trarre delle conclusioni.

Isomorfismo

Il punto finale circa il clustering gerarchico è l'idea di isomorfismo. Quando disegniamo un cluster, tutte le volte che abbiamo un nodo, ci si presenta una scelta: quale gene (o cluster) dobbiamo sistemare ed in quale lato del nodo? Pertanto, vi sono molti modi di disegnare lo stesso clustering, e perciò è importante ricordare che se appena due geni o cluster sono "vicini" l'un all'altro nel dendrogramma, questo non significa che siano vicini l'uno all'altro nel clustering. Bisogna sempre guardare l'albero per vedere le lunghezze dei rami in ordine a stabilire la prossimità.

Esempio 8.12: Cluster Isomorfici

Noi mostriamo due schemi dello stesso clustering dei cinque geni dal data set 7A (Figura 8.16). Benché il clustering sia lo stesso, i geni sono in ordine differente.

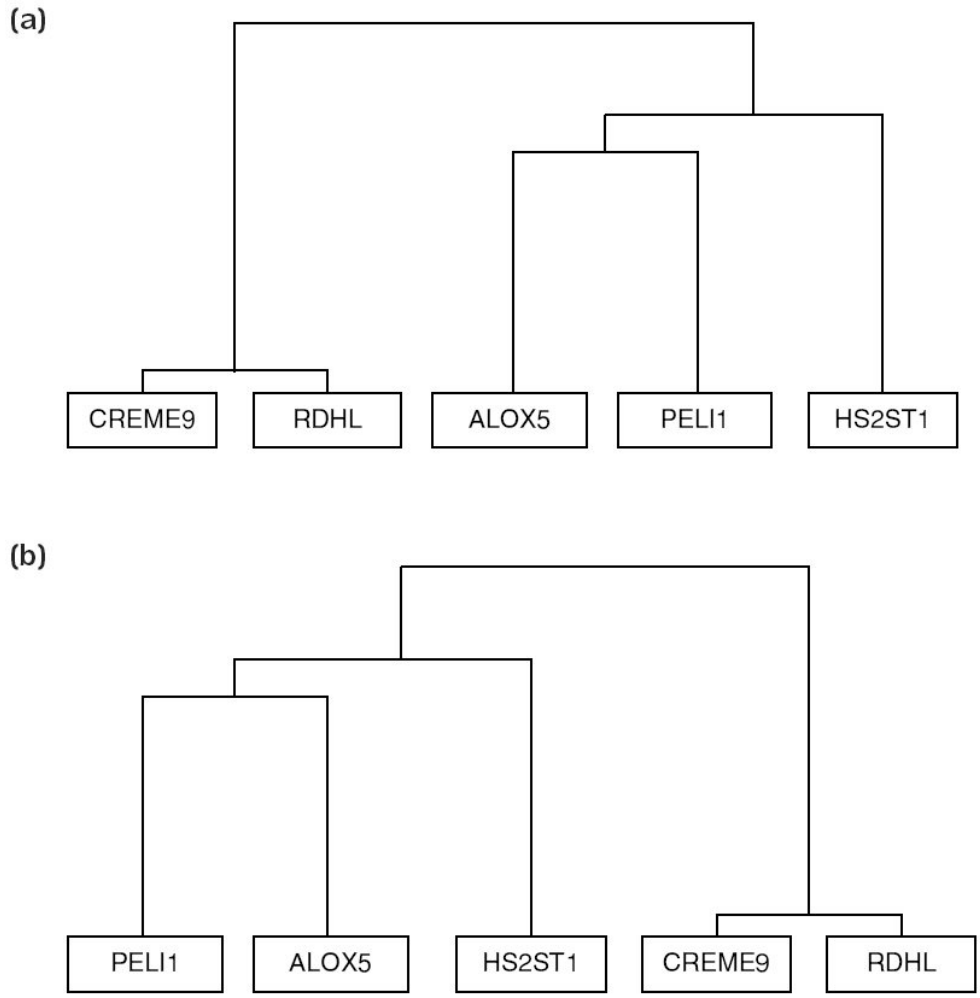


Figura 8.16: Isomorfismi. Isomorfismi. **(a)** Il dentrogramma per i cinque geni clusterati nell'Esempio 8.9. **(b)** Un dentrogramma alternativo per lo stesso clustering. Il clustering è identico, ma abbiamo disegnato il clustering in modo diverso.

8.5 L'Affidabilità e la robustezza del Clustering Gerarchico

Abbiamo descritto il clustering gerarchico - il metodo più comunemente usato per l'analisi dei microarray. Vi è una questione che ci si deve sempre porre in ciascuna analisi: quanto affidabili sono i risultati? I dati dei microarray sono affetti dal rumore, e spesso con elevati coefficienti di variazione. Abbiamo bisogno di essere sicuri che le strutture che vediamo nei digrammi del cluster rappresentino veramente gruppi correlati di geni e che non siano rappresentazioni del rumore random oppure errori sperimentali. Idealmente, si vorrebbe essere in grado di assegnare al cluster un fattore numerico di confidenza, in modo simile a quello con cui si assegna la deviazione o l'errore standard immediatamente a una misura numerica. Vi sono tre metodi con cui si assegna l'affidabilità in una analisi del cluster.

- **Visualmente.** Noi osserviamo i profili di espressione del gene nel cluster e vediamo se essi appaiano simili. Questo è un utile esercizio, ma interamente soggettivo ed inaffidabile, e diventa difficoltoso se il numero di geni e/o campioni diventa grande.
- **Attraverso la rilevanza biologica**. Se assumiamo che il cluster dei geni o dei campioni abbia un senso biologico, ci dovremmo aspettare che i geni rilevanti

biologicamente, oppure i campioni biologicamente simili, siano raggruppati nello stesso cluster. Per esempio, in Figura 8.12A, noi osserviamo che i geni che riparano il DNA sono stati annessi allo stesso cluster, e così i geni di risposta allo stress. Anche questo è un processo importante e utile, ma è ancora soggettivo, ed è facile pervenire a delle giustificazioni ad hoc a posteriori del perché particolari geni potrebbero essere raggruppati insieme senza considerare il data set come un tutt'uno e considerando i geni che non sono stati raggruppati in un cluster.

- **Uso di una misura statistica appropriata sulla conoscenza della variabilità sperimentale.** Sviluppando un tale genere di analisi, noi esprimiamo la relazione che intercorre tra la affidabilità del cluster e l'affidabilità dei dati sperimentali con cui essi (i cluster) sono costruiti. Dal punto di vista statistico, questo è il migliore approccio.

In questo paragrafo discutiamo una misura statistica con cui asserire quantitativamente l'affidabilità del cluster: La costruzione di un albero di consenso usando il bootstrapping parametrico. Il metodo degli alberi di consenso è ben chiaro e stabilito nell'analisi filogenetica ed è un eccellente metodo per l'analisi dei microarray.

Bootstrapping Parametrico

Noi abbiamo introdotto l'idea del bootstrapping nel Paragrafo 7.4. Lo spirito del bootstrap è quello di creare un data set immaginario che assomigli molto al data set originale. Per asserire l'affidabilità nell'analisi del cluster, noi vogliamo costruire il data set che potrebbe rappresentare un esperimento completamente separato ma identico a quello che abbiamo appena sviluppato. Qualsiasi differenza biologica tra i campioni o i trattamenti medici dovrebbe essere la stessa. Le sole differenze consistono nel fatto che i boostrap data set dovrebbero avere differenti errori come risultato, poiché provenienti da differenti esperimenti.

Noi applichiamo il bootstrap attraverso la conoscenza dei coefficienti di variabilità dell'esperimento[5], come misurato usando i metodi del Capitolo 6. Partiamo con i valori reali del rapporto logaritmico per ciascun gene. Quindi, costruiamo numeri randomici da una distribuzione normale con media zero e deviazione standard calcolata usando l'equazione 6.1, e quindi addizioniamo questi numeri randomici ai rapporti logaritmici reali. Il data set risultante è un boostrap data set: esso appare come l'originale data set, e gli errori che abbiamo addizionati ad esso sono esattamente della stessa ampiezza degli errori nel'esperimento.

Esempio 8.13: Bootstrap dati dal data set 8A

Noi costruiremo un Bootstrap data set per il gene MSH6 dal data set 8A. Il gene inizialmente non è differenzialmente espresso; diventa espresso alla quinta ora, raggiungendo un picco alla settima ora, e quindi inizia il decremento dell'espressione (Figura 8.13).

[5] La procedura che descriviamo é un bootstrap parametrico poiché essa costruisce i dati bootstrap addizionando deviazioni casuali da una distribuzione parametrizzata. In un esperimento con molti replicati, é anche possible applicare un bootstrap non parametrico. Invece di addizionare deviazioni random ai dati, noi selezioneremmo, per ciascuna misura per ciascun gene in ciascun campione, uno dei repliclati a caso ed useremmo quel valore come una misura dell'espressione. I bootstrap non parametrici hanno il vantaggio di permettere differenti livelli di errore per differenti geni e per una distribuzione degli errori non normale. Tuttavia, essi richiedono che l'esperimento sia ben replicato per avere successo. Il bootstrap parametrico ha il vantaggio che lo si puó applicare a qualsiasi esperimento, senza alcuna restrizione rispettto al numero di replicati usati

Il coefficiente di variabilitá di questo data set è di circa il 40%. Usando l'equazione 6.1, calcoliamo la deviazione standard degli errori normali dei valori logaritmici, che è 0.38.

Poiché stiamo usando i dati del rapporto logaritmico, dobbiamo moltiplicare questo risultato per $\sqrt{2}$, ed otteniamo una deviazione standard di 0.54. Questa è essenzialmente la deviazione standard per ciascun punto nella Figura 8.13.

Noi ora costruiamo i "dati bootstrap " per questo gene. Desideriamo che i "dati bootstrap " appaiano come i dati originali, con i punti temporali a 30 minuti, 2 ore, 5 ore, 7 ore, 9 ore ed 11 ore. Costruiamo sei variabili[6] randomiche ed addizioniamo queste ai dati reali (Tabella 8.7).

TABLE 8.7: Construction of Bootstrap Data for MSH6

Time	Real Data	Normal Random Variable	Bootstrap Data
30 minutes	−0.05	−0.10	−0.15
2 hours	−0.01	0.12	0.11
5 hours	1.75	−0.62	1.13
7 hours	3.80	−1.02	2.78
9 hours	3.28	0.43	3.71
11 hours	3.28	0.29	3.57

Note: Per costruire i dati di bootstrap si parte dai dati reali, quindi si costruiscono le variabili randomiche con la deviazione standard uguale alla variabilità misurata dei dati sperimentali, addizionando questi ai dati reali. Si noti che in questo esempio i dati reali raggiungono il picco dell'espressione dopo sette ore, ma i dati bootstrap raggiungono il picco dopo 9 ore. Ciò poiché la differenza tra i valori a 7 ore ed i valori a 9 ore è più piccola del livello di variabilità sperimentale, in tal modo non ci si può fidare della vera differenza dell'espressione del gene.

Ripetiamo questa procedura per ogni gene che analizziamo con l'analisi del cluster. Avendo costruito i dati bootstrap per ogni gene, sviluppiamo l'analisi del cluster su questi. Questo produce un nuovo dendrogramma che potrebbe essere differente dal dendrogramma originale (Figura 8.17).

[6] É immediato costruire numeri randomici normali in tutti i pacchetti statistici, come SPSS, R e SAS, ed é anche possibile fare questo con Excel usando l'add-in per l'analisi dei dati

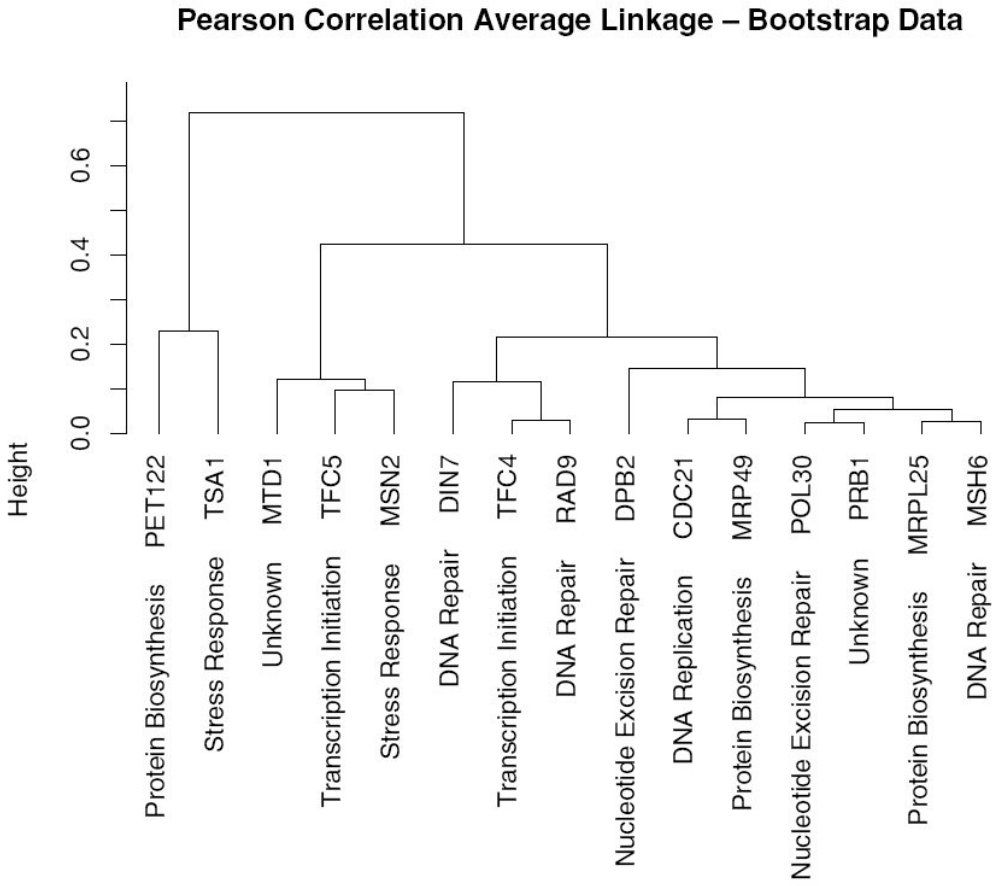


Figura 8.17: Clustering su un bootstrap data set. Deviazioni normali casuali vengono addizionate al "clustering data set" e i dati sono stati raggruppati di nuovo usando la correlazione di Pearson ed il linkage medio. Il clustering è simile al clustering nella Figura 8.14c: i dati appaiono separati (splitting) in geni persistenti e transitori, e molti dei geni che apparivano simili continuano ad essere racchiusi nello stesso cluster. Ma vi sono delle differenze, particolarmente nella struttura fine dei cluster. TCF5 si è anche spostato dal gruppo persistente al gruppo transitorio.

In questo esempio, il dendrogramma è simile al clustering dei dati reali: esso identifica i due cluster principali, cioè quello dei geni persistenti e quello dei geni transitori, e molti dei geni che sono vicini in Figura 8.14a lo sono anche in Figura 8.17. Ma vi sono alcune differenze; gran parte della struttura fine dei cluster è cambiata, ed il gene TFC5 appartiene ora al cluster dei geni transitori piuttosto che a quello dei geni persistenti.

Costruzione dell'Albero Consenso

Il bootstrap è applicato non una sola volta, ma parecchie volte. È abbastanza usuale creare almeno 1000 "bootstrap data set". Il prossimo è un passo importante. Noi siamo interessati ai cluster di geni che appaiono consistentemente nell'albero del bootstrap: questi sono i cluster che appaiono perfino quando addizioniamo il rumore ai dati che simulano errori sperimentali, e quindi essi sono cluster robusti agli errori sperimentali.

Ciascuna struttura del cluster che non appaia consistentemente negli alberi dei bootstrap non è robusta agli errori sperimentali, e vi è una certa difficoltà a tracciare l'inferenza statistica per tali tipi di cluster. I Matematici hanno dimostrato che i cluster che appaiono in piú del 50% degli alberi di bootstrap costituiscono un modello consistente, ed

è possibile costruire un albero da essi, denominato **albero consenso**. Non tutti i geni possono essere risolti con l'albero consenso: vi possono essere gruppi di geni che appartengono allo stesso cluster ma che non hanno alcuna struttura ulteriore. Questo non costituisce un problema; semplicemente significa che non vi è abbastanza evidenza sperimentale a suddividere quei geni in cluster più piccoli. Noi conosciamo anche il numero di volte che ciascun nodo sull'albero consenso è stato visto tra gli alberi del bootstrap. Ciò può essere usato come una misura della confidenza in ciascun nodo.

Pertanto, abbiamo raggiunto il nostro obiettivo - abbiamo identificato la struttura del cluster che è robusta agli errori sperimentali e, su ciascun nodo che rimane, abbiamo una quantificazione della confidenza in ciascun nodo dell'albero[7].

Esempio 8.14: Albero Consenso per il data set 8A

Noi facciamo girare (nel senso di far elaborare dal software) il bootstrap per 1000 volte per i 15 geni che abbiamo selezionato dal data set 9A, prima con la variabilità sperimentale misurata del 40%, e quindi con la variabilità del 30% per dimostrare la dipendenza dell'albero consenso dalla qualità dei dati.

Quando usiamo la variabilità del 40%, soltanto due cluster appaiono in più del 50% degli alberi: il cluster contenente gli 11 geni persistenti è apparso 590 volte, ed il cluster contenete i 2 geni transitori MSN2 e MTD1 è apparso 535 volte (Figura 8.18a). Nessuna altra struttura è stata osservata. I geni TSA1 e PET122 non sono associati in un cluster con nessun altro gene. Questo implica che la struttura fine vista nella Figura 8.12a non puó essere considerata con significato biologico: questo è tutto nell'ambito dei limiti degli errori sperimentali. Un fatto interessante, il gene TFC5 è stato annesso nel cluster con geni persistenti; il decremento dell'espressione del gene dopo 7 ore è dentro il 40% del'errore sperimentale e perciò esso non è significativo. Quando usiamo una variabilità del 30%, che è minore della variabilità osservata in questo particolare esperimento, ma potrebbe essere rappresentativo di un esperimento di migliore qualità, si osserva una piccola struttura. Vi è una separazione tra TFC5 e gli altri geni persistenti, mentre PET12 e TSA1 sono annessi ai cluster con MSN2 e MTD1 (Figura 8.18b). Non viene osservata nessun'altra struttura fine.

In sintesi, abbiamo imparato due lezioni dagli alberi consenso:

- Molta parte della struttura fine vista nell'analisi dei cluster può non essere significativa quando dobbiamo tenere in conto la variabilità sperimentale. L'albero consenso mostra la struttura che é affidabile.
- Migliore è il modo in cui si conduce l'esperimento, in termini di riduzione degli errori sperimentali e della variabilità, maggiore è l'informazione che si puó trarre dall'analisi del cluster.

[7] La costruzione dell'albero consenso é stata parte fondamentale di software filogenetici come Phylip per molti anni. Al tempo in cui questo libro é stato scritto, questi metodi non sono stati ancora inclusi in pacchetti software comunemente usato per l'analisi di microarray

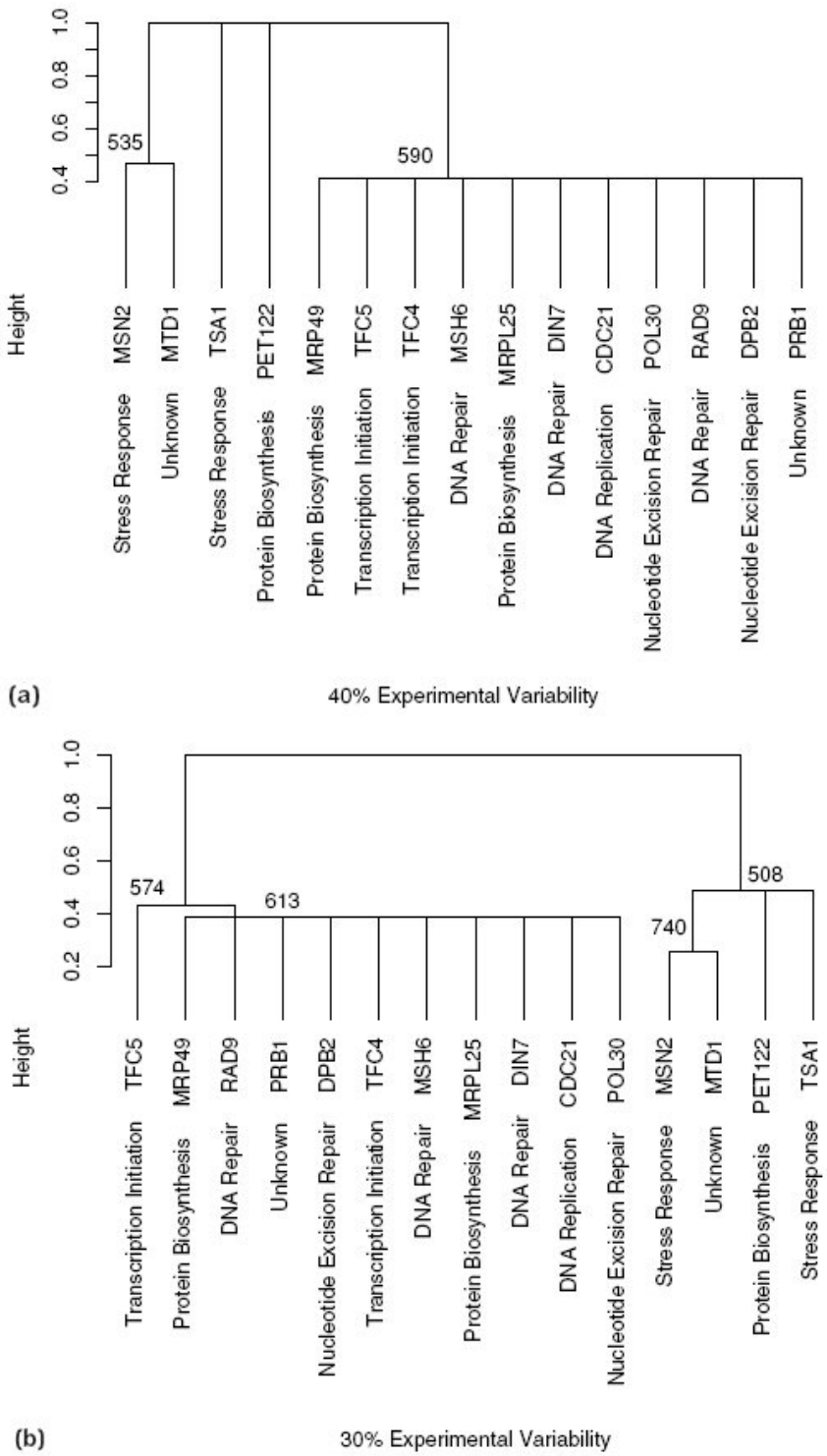


Figura 8.18: Costruzione dell'albero di consensi. Noi costruiamo gli alberi consenso per il clustering mostrato in Figura 8.14c usando 1000 replicati bootstrap. **(a)** Usiamo una variabilità sperimentale del 40%. Soltanto due cluster sono presenti nel'albero consenso. Il cluster contenente gli 11 geni persistenti è osservato in 590 dei 1000 alberi bootstrap, ed il cluster contenente MTD1 e MSN2 appare 535 volte. I geni PET122 e TSA1 non sono inclusi in un cluster. Nessun'altra struttura è osservata. Fatto interessante, il gene TFC5 è annesso al cluster dei geni persistenti: questo avviene poiché il decremento dell'espressione del gene dopo 7 ore è dentro il limiti dell'errore sperimentale. È difficile inferire qualsiasi altra struttura del cluster sulla base di questi dati, come la struttura vista in Figura 8.14c, poiché questa struttura è molto sensibile all'errore sperimentale. **(b)** L'albero consenso è costruito con un variabilità sperimentale del 30%. I cluster da (a) sono presenti con una evidenza più marcata; sono presenti anche molte strutture, con PET122 e TSA1, in clustering con MSN2 e MTD1, e TFC5 con leggera differenza dagli altri 10 geni persistenti. Concludiamo affermando che possiamo ottenere maggior informazione dai cluster da un migliore esperimento.

8.6 Machine-Learning Methods per l'analisi del cluster

Vi sono due metodi che sono stati messi a punto dalla comunità degli scienziati della "Learning Machine" che sono stati implementati in molti pacchetti software per l'analisi di espressione genica**: k-means clustering** e **self organized map**. Questo paragrafo fornisce una breve descrizione di tali metodi e mostra la loro applicazione ai data set che sono stati usati per il clustering gerarchico.

K-Means Clustering

K-Means è un algoritmo di clustering che differisce dal clustering gerarchico in tre punti essenziali:

- Il numero dei cluster deve essere pre-definito .
- Non esiste alcuna gerarchia o relazione tra i cluster, né esiste alcuna relazione o gerarchia tra i geni o i campioni al'interno del cluster; i cluster sono soltanto gruppi di profili similari di espressione dei geni.
- K-Means inizia con una allocazione randomica dei geni o dei campioni all'interno del cluster. Quindi, differenti runs (elaborazioni con il computer) di k-Means possono dare risultati leggermente diversi.

L'Algoritmo K-Means

L'algoritmo k-Means ha i seguenti sei passi:

- Scelta del numero dei cluster, indicato da k.
- Assegnazione randomica (casuale) di ciascun profilo di espressione dei geni ad uno dei k cluster.
- Calcolo del centroide di ciascuno dei k cluster.
- Per ciascun profilo, calcolo (un profilo alla volta) della distanza tra esso ed il centroide di ciascuno dei k cluster.
- Se il profilo è più vicino ad un cluster differente da uno in cui esso correntemente appartiene, spostare il profilo al nuovo cluster ed aggiornare i centroidi di entrambi i cluster.
- Tornare indietro al passo 4 e ripetere fino a quando nessun profilo cambia la classe di appartenenza.

Esempio 8.15: K-Means Clustering del data set 8A

Applichiamo il k-Means clustering ai 15 geni ai quali abbiamo applicato il clustering gerarchico usando la correlazione come misura della distanza. I risultati per 2, 3, 4 e 5 cluster sono mostrati in Tabella 8.8. Questo clustering trova anche i gruppi di geni persistenti e transitori. Quando K = 2, il gene POL30 è raggruppato con i geni transitori.

Questo è probabilmente un risultato anomalo poiché vi sono troppo pochi cluster. Il valore di k = 4 sembra fornire un buon risultato. Quando K = 5, i geni persistenti si separano in due gruppi e i dati sono probabilmente divisi in un numero eccessivo di cluster.

TABLE 8.8: K-Means Clustering of Genes from Data Set 8A

Number of Clusters (k)			
$k=2$	$k=3$	$k=4$	$k=5$
Cluster 1	**Cluster 1**	**Cluster 1**	**Cluster 1**
MTD1	MTD1	MTD1	MTD1
MSN2	MSN2	MSN2	MSN2
TSA1	TSA1	TSA1	TSA1
POL30	**Cluster 2**	TFC5	TFC5
Cluster 2	POL30	**Cluster 2**	**Cluster 2**
PET122	**Cluster 3**	PET122	PET122
TFC5	PET122	**Cluster 3**	**Cluster 3**
DIN7	TFC5	POL30	POL30
RAD9	DIN7	**Cluster 4**	**Cluster 4**
PRB1	RAD9	DIN7	DIN7
CDC21	PRB1	RAD9	RAD9
MSH6	CDC21	PRB1	PRB1
MRP49	MSH6	CDC21	**Cluster 5**
MRPL25	MRP49	MSH6	CDC21
TFC4	MRPL25	MRP49	MSH6
DPB2	TFC4	MRPL25	MRP49
	DPB2	TFC4	MRPL25
		DPB2	TFC4
			DPB2

Note: Applichiamo il k-means clustering ai 15 geni studiati, variando il numero dei clusters k da 2 a 5. L'algoritmo di clustering trova i gruppi persistenti e transitori. Quando k=2, POL30 è clusterato con i geni transitori, probabilmente perché vi sono troppo pochi clusters. Quando k=4, la struttura del cluster sembra interpolare bene i dati. Quando k=5, i geni persistenti si separano in due gruppi; i dati sono stati probabilmente over-clusterati.

Come scegliere un buon valore di *K*?

Il numero di cluster deve essere scelto a priori. Ciò significa che l'operatore deve fare qualche tentativo per stimare il numero di cluster prima di far girare (elaborare) l'algoritmo. Vi sono due approcci per la stima del numero di cluster da usare: per mezzo di MDS ed empiricamente. Noi raccomandiamo di usare entrambi.

MDS (Paragrafo 8.3) permette di visualizzare la distanza tra i geni od i campioni nello spazio bidimensionale. Adottando il metodo di misura della distanza che si intende usare per il clustering, si può pervenire ad una indicazione se vi sia un numero naturale di cluster nei dati.

Esempio 8.16: MDS per aiutare a scegliere *k*

Noi applichiamo MDS a 15 geni (Figura 8.19). Vi è un cluster principale di geni sulla parte sinistra dello spazio; questo potrebbe essere dovuto a due cluster con MSH6, MRPL25 e TCF4 lievemente lontani dal gruppo principale. MTD1 e MSN2 sono vicini insieme e potrebbero formare un altro cluster; TFC5 sembra che sia localizzato tra questi ed il cluster grande. DIN7, PET122 e TSA1 sono lontani da altri geni. Quindi, noi avremmo necessità di tre cluster come minimo, e probabilmente di 6 cluster al massimo.

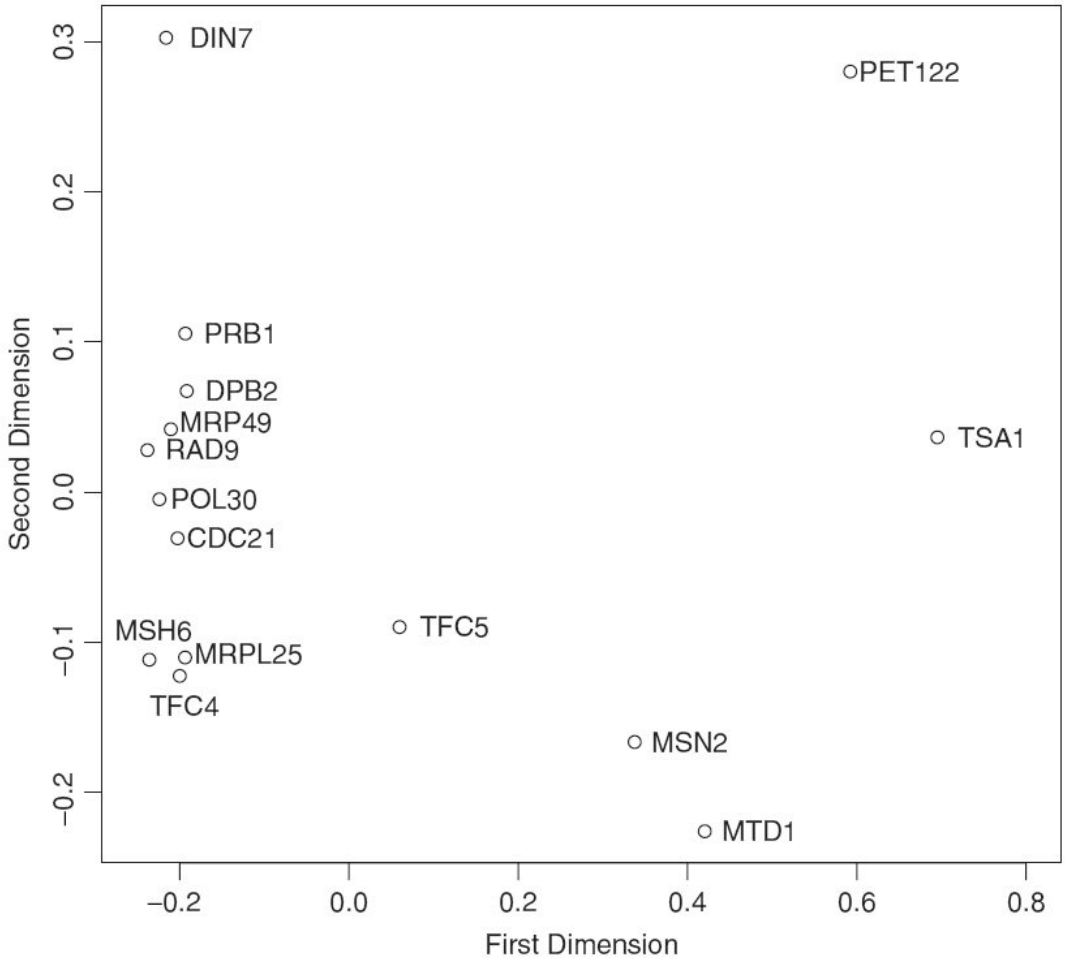


Figura 8.19: Scaling Multidimensionale e Clustering dei dati. Noi applichiamo MDS a 15 geni che stiamo usando per il clustering, facendo uso della correlazione di Pearson come misure della distanza. Vi è un grande cluster di geni alla parte sinistra della figura, che é un sottocluster di MSH6, MRPL25 e TFC4. I geni MSN2 e MTD1 stano tutti vicini, e TCF5 si trova tra questi ed il cluster principale. DIN7, PET122 e TSA1 appaiono essere fuori posto. Noi dovremmo essere in grado di stimare, da questa figura, che vi potrebbero essere tra quattro e sei cluster naturali in questi dati.

Una scelta empirica di k significa far girare l'algoritmo k-means, con differenti numeri di cluster, e differente metrica della distanza, con testing di affidabilità dei cluster usando uno (o tutti) dei differenti metodi descritti. Noi potremmo quindi selezionare un numero di cluster che fornisce un risultato robusto, affidabile e significativo.

Validazione K-Means Clustering

Come con il clustering gerarchico, è essenziale validare i risultati del clustering k-means. Vi sono quattro modi per validare i risultati; questi sono simili ai metodi descritti nel Paragrafo 7.4

- **Visivamente**. Guardare per vedere se geni o campioni nello stesso cluster abbiano profili similari
- **validità Biologica**. Guardare per vedere se i geni nello stesso cluster abbiano funzioni biologiche simili o complementari, oppure se i campioni nello stesso cluster siano derivati da sorgenti biologiche simili.

- **Ripetendo il clustering** i dati con lo stesso valore di k. Vi è un elemento randomico nel clustering k-means. Se vengono fuori gli stessi cluster è segno che il clustering sta funzionando bene. Se ciascuna operazione di clustering dà cluster differenti, può darsi che si stia usando un valore di k sbagliato, una misura di distanza sbagliata, oppure puó essere che questo non sia un buon metodo di analisi dei dati.
- **Analisi statistica.** Il bootstrapping parametrico (Figura 8.5) può essere applicato anche al clustering k-means. Si dovrebbe credere soltanto ai cluster che appaiono un maggior numero di volte nel bootstrap clustering, ed il bootstrap dovrebbe permettere di impostare una misura di confidenza su quei cluster.

Mappe Auto-Organizzate

L'ultimo algoritmo di clustering che vedremo in questo capitolo sono le mappe auto-organizzate. Questi metodi di clustering sono simili al k-means in quanto l'utilizzatore specifica un numero predefinito di cluster. Tuttavia, ad eccezione di k-means, i cluster sono correlati l'uno all'altro attraverso una topologia spaziale. Di solito, i cluster sono distribuiti come una griglia rettangolare. L'algoritmo è abbastanza dettagliato e rinviamo il lettore interessato ai riferimenti alla fine del capitolo. Vi sono tre importanti proprietà delle mappe auto-organizzate:

- I cluster sono correlati l'uno all'altro secondo una topologia spaziale, di solito in una griglia.
- La grandezza della griglia (numero di cluster) deve essere scelta a priori; di solito è senso comune provare diverse grandezze e vedere quale funziona meglio.
- I geni sono allocati in un primo momento dentro i cluster in modo randomico, così che differenti runs (runs del software) della mappa auto organizzata possono dare differenti risultati.

Esempio 8.17: Mappe auto-organizzate sul data set 8A

Facciamo girare la mappa auto-organizzata con una griglia 2 x 2 degli stessi geni che abbiamo usato per tutti i metodi di clustering. Mostriamo i profili medi di quattro cluster della Figura 8.20. I quattro cluster hanno una relazione spaziale: i due cluster in alto rappresentano le risposte persistenti, mentre i due cluster in basso sono le risposte ai transienti. In questo caso, i due cluster in alto sono molto simili, probabilmente perché la griglia 2 x 2 contiene parecchi cluster per questo data set.

In tabella 8.9 mostriamo l'allocazione dei geni ottenuta con applicazioni distinte della mappa auto-organizzata. L'allocazione dei geni nei cluster è differente. Questo è un problema serio quando si usino le mappe auto-organizzate per l'analisi di dati per i microarrays, e deve essere tenuto in considerazione tutte le volte che si usano le mappe auto-organizzate.

Scelta della grandezza delle mappe auto-organizzate

Per la determinazione della grandezza della mappa auto-organizzata, dovrebbe essere applicato esattamente lo stesso principio che si segue per scegliere il numero di cluster *k*.

Si può usare MDS per vedere se vi è un numero naturale di cluster, e si può scegliere la grandezza della mappa empiricamente.

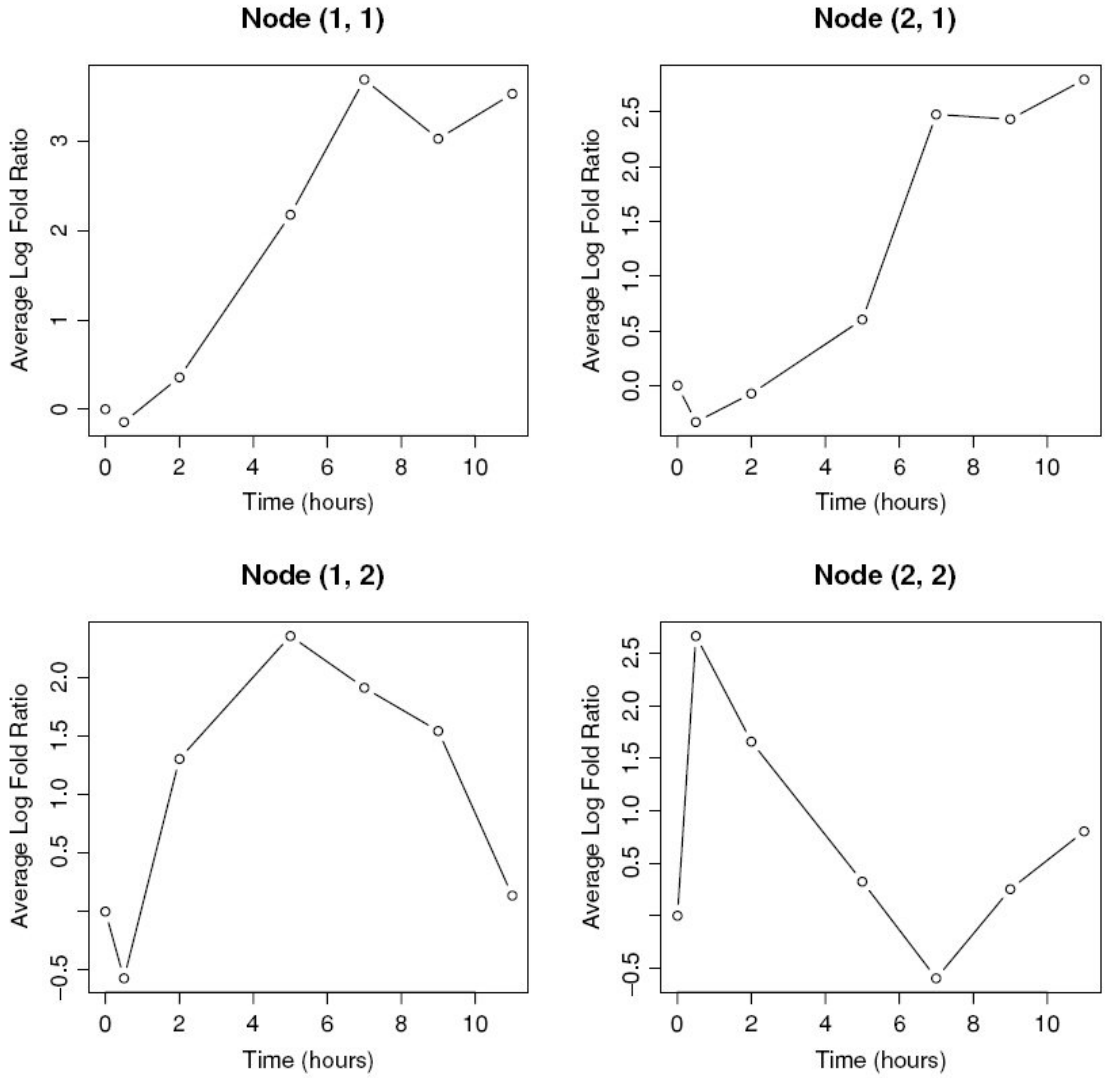


Figura 8.20: Profilo di geni per quattro nodi della mappa auto-organizzata. Ciascuno dei quattro nodi della mappa auto-organizzata ha profili medi mostrati nei quattro pannelli della figura. I due nodi in alto sono quasi simili; entrambi contengono geni persistentemente regolati. I due nodi in basso contengono geni regolati transitoriamente.

Esempio 8.18: Scelta della grandezza della mappa *Auto-Organizzata*

Guardando ancora al diagramma MDS (Figura 8.19), possiamo vedere che quattro cluster dovrebbe essere un numero ragionevole da usare, ma i 9 cluster forniti da una griglia 3 x 3, potrebbero essere troppi cluster per questi dati. Pertanto, nell'esempio 8.17 usiamo una griglia 2 x 2 per ottenere quattro cluster.

TABLE 8.9A: Self-Organised Maps Applied to Data Set 8A

CDC21	DIN7
MSH6	RAD9
MRP49	PRB1
MRPL25	DPB2
TFC4	
TFC5	
MTD1	POL30
MSN2	PET122
TSA1	

Note: Allocazioni dei geni al cluster dall'elaborazione 2x2 della mappa auto-organizzata che é stata usata per creare la figura 8.16

TABLE 8.9B: Self-Organised Maps Applied to Data Set 8A

CDC21	PET122
MSH6	
MRP49	
MRPL25	
TFC4	
TFC5	
DIN7	
RAD9	
PRB1	
DPB2	
MTD1	POL30
MSN2	TSA1

Note: Allocazioni dei geni ai clusters da una elaborazione separate di una mappa auto-organizzata 2x2. Si noti che le allocazioni sono differenti nelle due elaborazioni. Vi è un elemento random nella mappa auto-organizzata, che significa che due elaborazioni possono essere molto differenti.

Validazione delle Mappe Auto-organizzate

Così come con il clustering k-means, le mappe auto-organizzate possono essere validate con quattro metodi differenti:

- **Visivamente.** Guardando per vedere se vi siano profili simili all'interno del cluster e se ci siano differenze tra i cluster.
- **Rilevanza biologica**. Controllare che i geni con funzione biologica simile, oppure campioni provenienti da sorgenti biologiche simili, siano presenti nello stesso, oppure in cluster vicini.
- **Reclustering**. La mappa auto-organizzata cambia vistosamente quando si fa il re-run della mappa stessa? Se così fosse, il cluster che cambia può essere difficoltoso ad essere interpretato.
- **Analisi Statistica**. Benché sia possibile usare i metodi di bootstrap descritti nel paragrafo 8.5 alle mappe auto-organizzate, i cluster consenso non sono correlati l'uno all'altro in accordo a qualche topologia spaziale, e quindi la topologia originale delle mappe auto-organizzate può essere perduta.

TABLE 8.10: Advantages and Disadvantages of Different Clustering Methods

Hierarchical	Hierarchical Consensus Tree	K-Means	Self-Organised Maps
✓ Easy to understand algorithm	✓ Intuitive interpretation of results	✓ Intuitive interpretation of results	✓ Implemented in many gene expression packages
✓ Intuitive interpretation of results	✓ Robust to noise and errors	✓ No preimposed hierarchy	× User has to specify number of clusters in advance
✓ Same data give the same results every time	✓ Hierarchy is representative of statistically significant differences in gene expression profiles	✓ Implemented in many gene expression packages	× No a priori reason why gene expression clusters should fit a two-dimensional topology
✓ Implemented in many gene expression packages	✓ Provides a measure of confidence in clusters	× User has to specify number of clusters in advance	× Can get very different results when run different times
× No a priori reason why genes should relate on a binary tree	× Not yet implemented in gene expression analysis software	× Can give different results when run different times	× Cannot easily construct consensus clusters without losing topology
× Noise and errors can adversely influence a tree	× Can give slightly different answers on different runs	× No measure of confidence in results although can construct consensus clusters	
× No measure of confidence in results			

Riassunto dei punti chiave

Nella Tabella 8.10, abbiamo riassunto i punti chiave dei vantaggi e degli svantaggi di differenti metodi di clustering. I punti chiave da considerare in questo capitolo sono i seguenti:

- Vi è un certo numero di modi per misurare la similarità tra due profili di espressione del gene, e la misura che si utilizza perturberà i risultati. Quindi si raccomanda di far girare l'analisi con un numero differente di misura delle distanze.
- PCA e MDS forniscono un buon modo di visualizzare i dati senza imporre che questi abbiano alcuna gerarchia.
- Il clustering gerarchico puó essere usato per identificare geni correlati oppure campioni, e disegnare questi usando un dendrogramma.
- Vi sono molte varianti del clustering gerarchico, ciascuna delle quali puó produrre risultati differenti. Si dovrebbero sempre provare differenti metodi di linkage e metrica della distanza.
- I metodi Machine Learning possono anche essere usati per definire la relazione tra geni o campioni, ma possono produrre differenti risultati ogni volta che si fanno girare.
- Si dovrebbe sempre sviluppare una validazione statistica sui risultati (e.g., usando un algoritmo di bootstrap).

Capitolo 9

Classificazione dei Tessuti e dei Campioni

9.1 Introduzione

Una delle aree più eccitanti della ricerca con i microarrays, è il loro utilizzo per trovare gruppi di geni che possano essere usati in diagnostica per determinare il disturbo di cui un individuo soffre, oppure per predire - dal punto di vista della prognosi - il successo o meno, di una sessione terapeutica, oppure per predire i risultati di un esperimento.

In questi studi, i campioni sono prelevati da parecchi gruppi di individui con patologie, sintomi o fenotipi conosciuti, e sono ibridizzati ai microarrays. Lo scopo è quello di trovare un piccolo numero di geni che possano predire a quale gruppo ciascun individuo appartenga. Questi geni possono essere usati in futuro su ulteriori individui come parte di un test molecolare, sia usando un microarray focalizzato allo scopo, sia un metodo più semplice come, per esempio, una reazione a catena della polimerasi (PCR).

Esempio 9.1: ATA SET 9A

Campioni di midollo osseo sono prelevati da 27 pazienti affetti da leucemia linfoblastica acuta (ALL) e da 11 pazienti affetti da leucemia mieloide acuta (AML) ed ibridizzati agli arrays Affymetrix[1]. Lo scopo che ci prefiggiamo è di essere in grado di diagnosticare la leucemia in futuri pazienti usando sia la tecnologia Affymetrix , sia arrays più specifici con un piccolo numero di geni. Come dobbiamo scegliere una serie di regole per classificare questi campioni?

Lo sviluppo dei modelli predittivi dipende dalla statistica e dalle tecniche computazionali, molte delle quali sono tutt'ora oggetto di una attiva investigazione. Vi sono essenzialmente tre filoni nello sviluppo di modelli predittivi, e dunque abbiamo strutturato il capitolo nei tre successivi paragrafi:

Paragrafo 9.2: Metodi di Classificazione, si occupa un insieme di metodi usati più comunemente per distinguere tra gruppi o individui sulla base di un data set di misure. Vi sono parecchi metodi ben accertati per far ciò, parecchi dei quali funzionano particolarmente bene con i dati di un microarray.

Paragrafo 9.3: Validazione, investiga quei metodi adatti a verificare i risultati di una analisi di classificazione. Esso discute i due approcci usati piú comunemente: **training e test sets**, e **cross validazione (validazione incrociata).**

Paragrafo 9.4: Riduzione della dimensionalitá, investiga un insieme di metodi che possono essere usati per trovare un appropriato e piccolo insieme di geni - partendo da un grande insieme di geni - su un microarray per distinguere tra i gruppi di individui in una maniera robusta ed affidabile. Questo è un aspetto della ricerca ancora aperto; noi descriviamo solo alcuni di questi metodi - quelli che sono già in uso, ma a tutt'oggi mancano standards ben stabiliti.

[1] I dati sono stati presi dall'articolo di Golub ed altri (1999). Il riferimento é fornito alla fine del capitolo. I dati sono disponibili allo Stanford Microarray Database.

9.2 Metodi di Classificazione

In questo paragrafo descriviamo i metodi che permettono di predire la classe a cui un individuo appartiene, basandosi sulle misure di espressione del gene. Per far ciò, costruiamo un modello predittivo usando le misure di espressione del gene di individui dei quali conosciamo la classe di appartenenza. Nel linguaggio della Machine Learning Community, ciò è conosciuto con il nome di **apprendimento supervisionato**. Nel corso di questo paragrafo, assumiamo sempre di aver selezionato un piccolo numero di geni dei quali usiamo le misure di espressione, invece di usare tutti i geni sul microarrays. Nel paragrafo 9.4 discutiamo i metodi che useremo per selezionare i geni che ci permettono di costruire buoni modelli predittivi.

Iniziamo analizzando due concetti che sono fondamentali per la classificazione: la **separabilità** e la **linearità**. Quindi descriveremo cinque differenti metodi di classificazione: **k-nearest** (elemento piú vicino), **centroide piú vicino**, **analisi del discriminante lineare**, **reti neurali** e **macchine di supporto vettoriali**. Questi sono tutti metodi che sono stati applicati per l'analisi dei dati dei microarrays.

Separabilità

Supponiamo di avere le misure di espressione di un gene per un certo numero di campioni. Nelle analisi di classificazione, é utile pensare a ciascun campione come se occupasse una locazione in uno spazio a molte dimensioni. Ciascun asse nello spazio è la misura dell'espressione di uno dei geni. Se noi stiamo usando le misure di due geni, allora il campione può essere immaginato come se avesse una locazione nello spazio bidimensionale; se stiamo usando le misure di tre geni, allora dovremmo pensare ad un campione come occupante una locazione nello spazio tridimensionale; e se stiamo usando le misure di dieci geni, dovremmo pensare ad un campione come occupante una locazione nello spazio a dieci dimensioni.

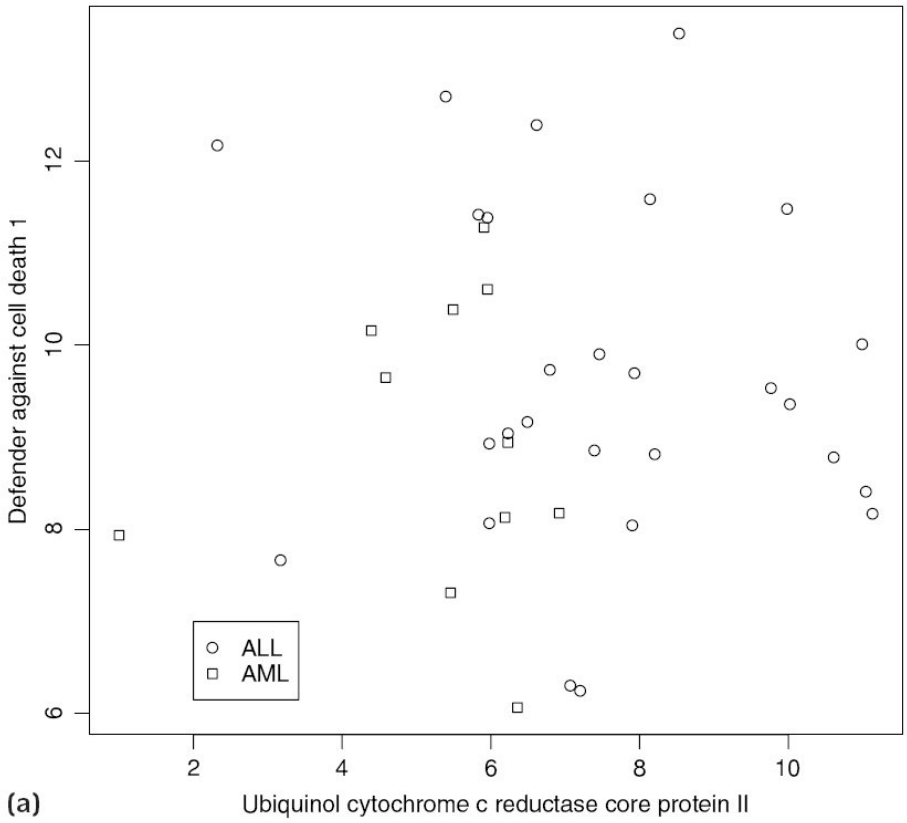
Esempio 9.2: concettualizzazione delle misure del campione nello spazio di espressione del gene

Nel data set 8A, vi sono 27 pazienti ALL ed 11 pazienti AML. Se noi consideriamo solo il gene *ubiquinolo citocromo c reduttasi proteina core II* ed il *difensore contro la morte cellulare I* allora possiamo pensare che ciascun campione abbia una posizione nello spazio bidimensionale, essendo ciascun asse il livello di espressione di ciascuno dei due geni (Figura 9.1a).

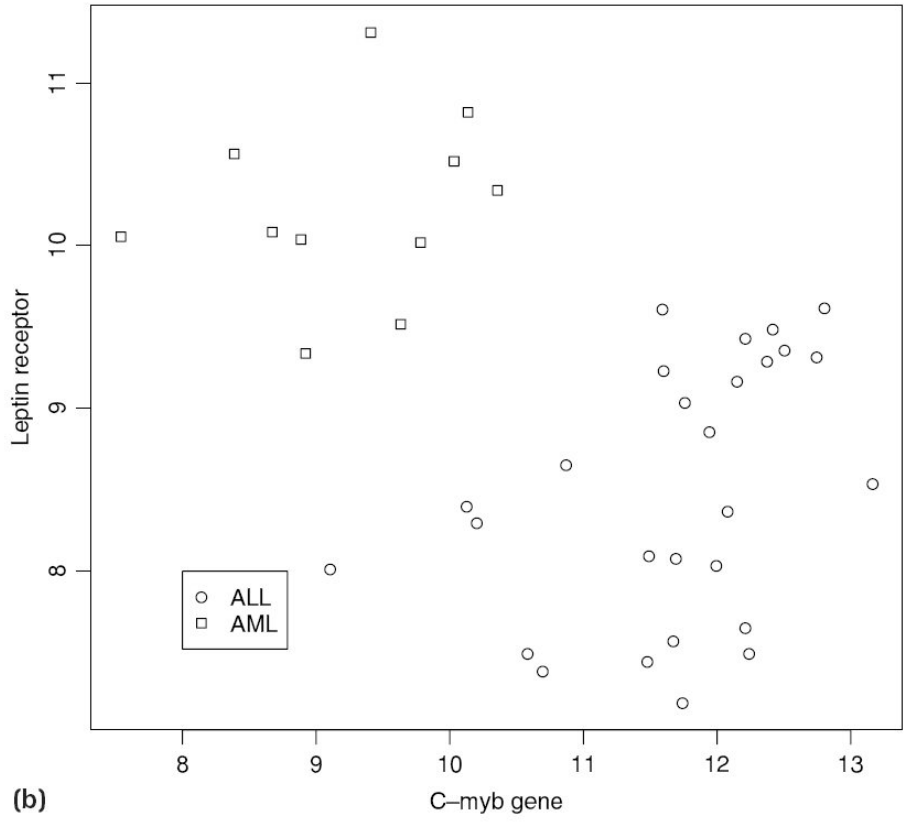
Con questo framework (assetto di lavoro), vi sono due scenari estremi:

- **Separabile**: I differenti gruppi a cui i campioni appartengono occupano differenti regioni dello spazio di espressione del gene.
- **Non separabile**: I differenti gruppi a cui i campioni appartengono sono miscelati insieme nella stessa regione dello spazio di espressione del gene.

In molti casi, i dati possono essere solo parzialmente separabili, con differenti gruppi largamente distribuiti ed occupanti differenti regioni dello spazio ma con qualche sovrapposizione nelle zone di confine. Lo spirito dei metodi di classificazione descritti in questo paragrafo è quello di trovare un modo di partizionare lo spazio così che ciascun gruppo sia in una differente regione e di descrivere le partizioni, così che si possa determinare a quale gruppo appartenga un nuovo campione. Come parte di questo processo, noi possiamo quantificare l'entità con cui il dato è separabile.


Defender against cell death 1
6
8
10
12
2
4
6
8
10
ALL
AML
Ubiquinol cytochrome c reductase core protein II

(a)


Leptin receptor
8
9
10
11
8
9
10
11
12
13
ALL
AML
C–myb gene

(b)

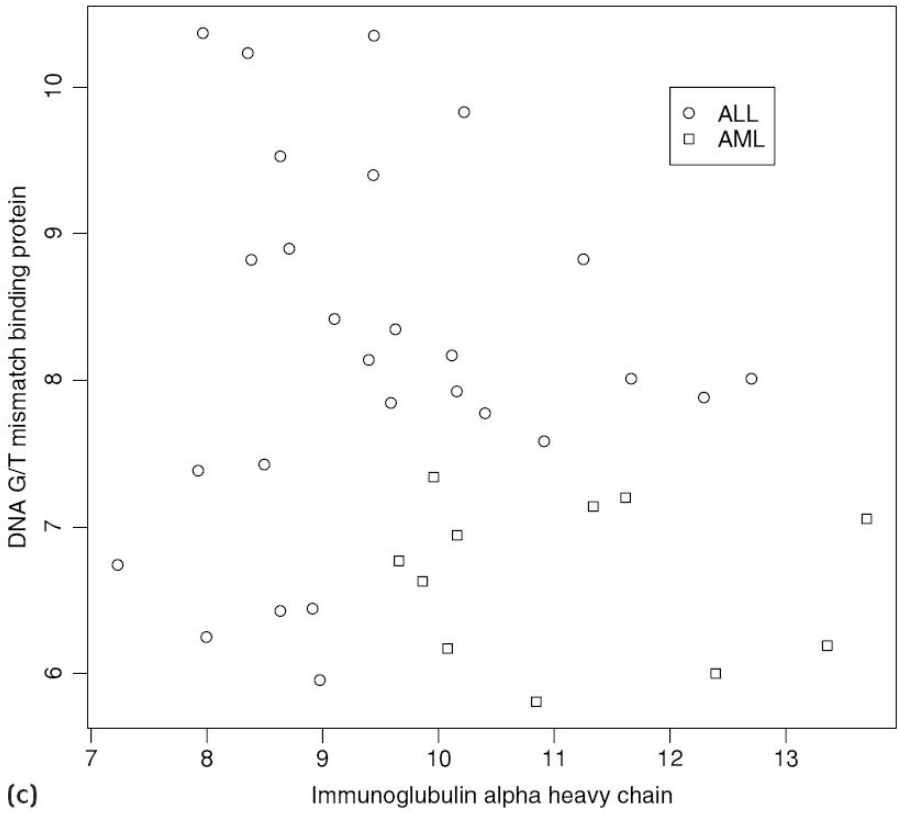


Figura 9.1: Separabilità delle classi. Ventisette campioni provenienti da pazienti affetti da ALL ed 11 campioni provenienti da pazienti affetti da AML sono stati ibridizzati ai microarrays (data set 9A). Noi vorremmo trovare i gruppi di geni che possano essere distinti tra i due gruppi. Per scopi illustrativi, mostriamo gli esempi con soltanto due geni. In generale, è comune usare più di due geni per le analisi di classificazione. **(a)** I geni ubiquinolo citocromo c reduttasi proteina core II e difensore contro la morte cellulare I non possono essere distinti tra i due gruppi. Noi denominiamo i dati che si comportano in questo modo *inseparabili*. **(b)** i geni C-myb gene ed recettore della leptina dividono i dati in due gruppi. In questo caso, sarebbe possibile tracciare una linea retta tra le classi ALL ed AML per separare i gruppi: noi chiamiamo i dati *linearmente separabili*. Metodi come l'analisi del discriminante lineare dovrebbero funzionare bene con questi dati. **(c)** Il gene per la catena pesante alfa delle immunoglobuline e il gene per la proteina che lega il mismatch G/T nel DNA separano anche essi le due classi. Ma in questo caso, non vi è una linea retta che possa essere usata per separare i due gruppi. Noi necessiteremmo di due linee rette, o di due linee curve per distinguere le classi. Denominiamo questi dati *non linearmente separabili*.

In un esperimento con i microarrays, il livello di separabilità dei dati sarà determinato dal gruppo di geni di cui stiamo considerando le misure di espressione. Parte del compito dei metodi di riduzione della dimensionalità dei dati descritti nel paragrafo 9.4 è di trovare i gruppi di geni che massimizzano la separabilità dei dati.

Esempio 9.3: Separabilità e non separabilità dei dati

Se noi consideriamo la misura del campione nel data set 9A nello spazio a due dimensioni con i geni ubiquinolo citocromo c reduttasi proteina core II ed il difensore contro la morte cellulare I, allora i due gruppi di pazienti (ALL ed AML) non occupano differenti regioni dello spazio (Figura 9.1a). I dati non sono separabili.

Se noi consideriamo gli stessi pazienti ma usiamo le misure dei geni C-myb e recettore della leptina, allora i due gruppi sono in regioni differenti dello spazio (Figura 9.1b). I dati sono separabili, ed è possibile costruire un classificatore usando questi geni.

Linearità

Quando consideriamo dati separabili, vi sono due possibilità riguardanti il modo in cui i campioni possono essere sparati nello spazio:

- **Linearmente separabile**. I dati sono linearmente separabili se è possibile operare una partizione dello spazio tra i due (o più) gruppi usando linee rette.
- **Non-Linearmente separabile**. I dati sono non linearmente separabili se i due gruppi sono separabili, ma non è possibile una partizione tra i gruppi usando linee rette.

Descriveremo alcuni metodi che si applicano soltanto alle tecniche di separazione lineare, ed altri metodi che sono adatti a classificare i dati separabili non linearmente. I dati di microarray sono frequentemente non lineari, e quindi spesso viene raccomandato di provare sia la metodologia lineare che quella non lineare.

Esempio 9.4: Dati separabili linearmente e non linearmente

Il data set 9A puó essere linearmente e non linearmente separabile, in dipendenza di quali misure di espressione del gene vengono considerate. Con i geni C-myb e recettore della leptina, i dati sono linearmente separabili (Figura 91b). I geni per la catena pesante alfa delle immunoglobuline ed per la proteina che lega il missmatch G/T nel DNA appaiono ancora separati in due gruppi, ma la separazione non è lineare (Figura 9.1c).

Numero di Classi

Nel corso di questo capitolo usiamo il data set 9A, che consiste di due classi, AML ed ALL. La ragione di ciò è che esso è il metodo di classificazione più facile da comprendere dal punto di vista del "*distinguere*" tra due gruppi. Di converso, molte applicazioni di classificazione che usano i microarrays possono avere più di due gruppi di individui. Tre dei cinque metodi che descriveremo si estendono molto naturalmente a questi dati.

Esempio 9.5: Dati di microarray con 4 classi: data set 9B

Vi sono quattro tipi di tumori dell'infanzia, caratterizzati da piccole cellule blu rotonde: neuroblastoma (NB), linfoma non Hodgkin (NHL), rhabdomyosarcoma (RMS) e tumori di Ewing (EWS). Sessantatre campioni provenienti da questi tumori sono stati ibridizzati ai microarray[2]. Noi desideriamo essere in grado di diagnosticare questi tumori nei bambini usando microarrays focalizzati allo scopo ed un set di regole per distinguere i tipi di tumore.

Metodi di Classificazione

[2] I dati provengono dall'articolo di Khan e al. (2001). Il riferimento e l'URL per i dati sono forniti alla fine del capitolo

Descriviamo cinque differenti metodi per la partizione dello spazio e la predizione del gruppo in un nuovo campione:

- K-nearest neighbours (elementi piú vicini)
- Classificazione dei centroidi
- Analisi del discriminante lineare
- Reti neurali
- Macchine di supporto vettoriali

Questi sono metodi ben stabiliti e di uso comune; essi sono implementati in molti pacchetti specializzati per l'analisi dei dati, come per esempio **R** o **Mathlab**. I vantaggi e gli svantaggi di questi metodi sono riassunti nella Tabella 9.1.

TABLE 9.1: Advantages and Disadvantages of Five Classification Algorithms

K-Nearest Neighbours	Centroid Classification	Linear Discriminant Analysis	Neural Networks	Support Vector Machines
✓ Intuitive and easy to understand	✓ Intuitive and easy to understand	✓ Strong statistical theory	✓ Able to discriminate non-linearly separable data	✓ Able to discriminate non-linearly separable data
✓ Extends naturally to more than two classes	✓ Extends naturally to more than two classes	✓ Generally outperforms centroid classification	✓ Extends naturally to more than two classes	✓ Faster to train than neural networks
✓ Separates non-linearly separable data	✓ It is very fast: there is no training time	× Does not extend naturally to more than two classes	× Slow to train	× Does not extend naturally to more than two classes
✓ It is very fast: there is no training time	× Gives incorrect results on non-linear data	× Gives incorrect results on non-linear data	× Have to optimise architecture	× Have to optimise kernel function
× Not robust to outliers				

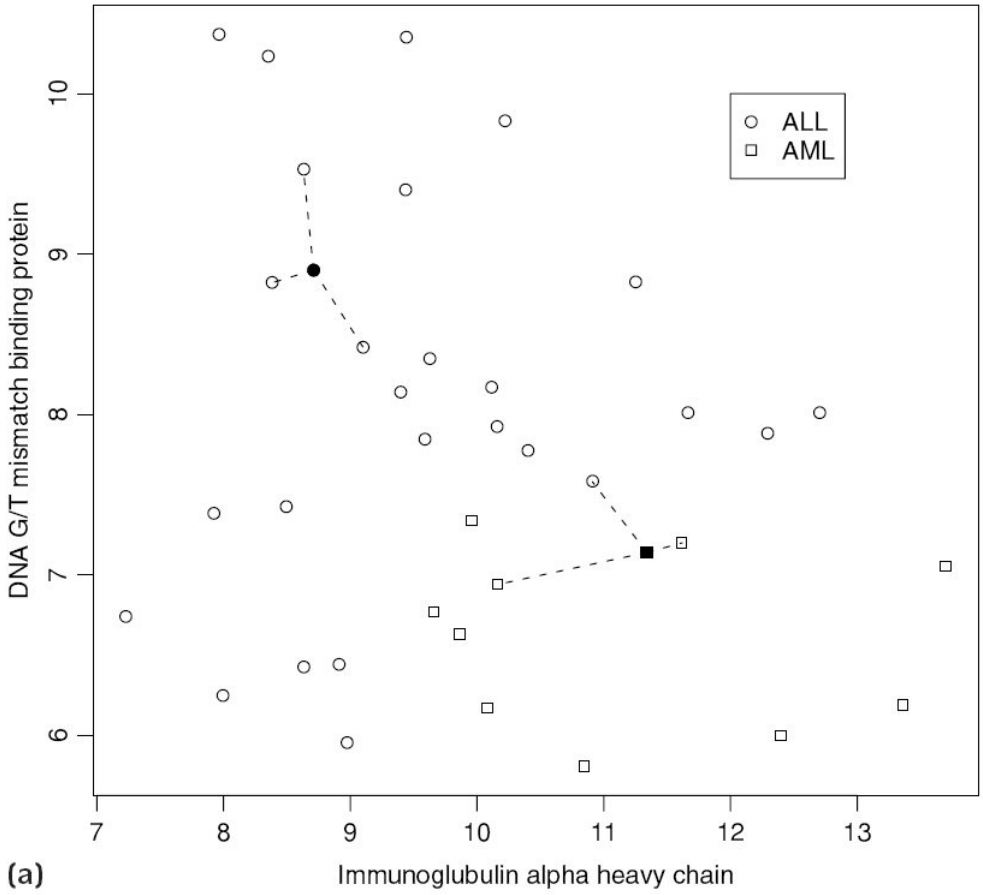
K-Nearest Neighbours

K-Nearest Neighbours (KNN) è il metodo più semplice per decidere la classe a cui il campione appartiene (Figura 9.2). Abbiamo un certo numero di campioni la cui classe di appartenenza è nota. Vi sono tre passi:

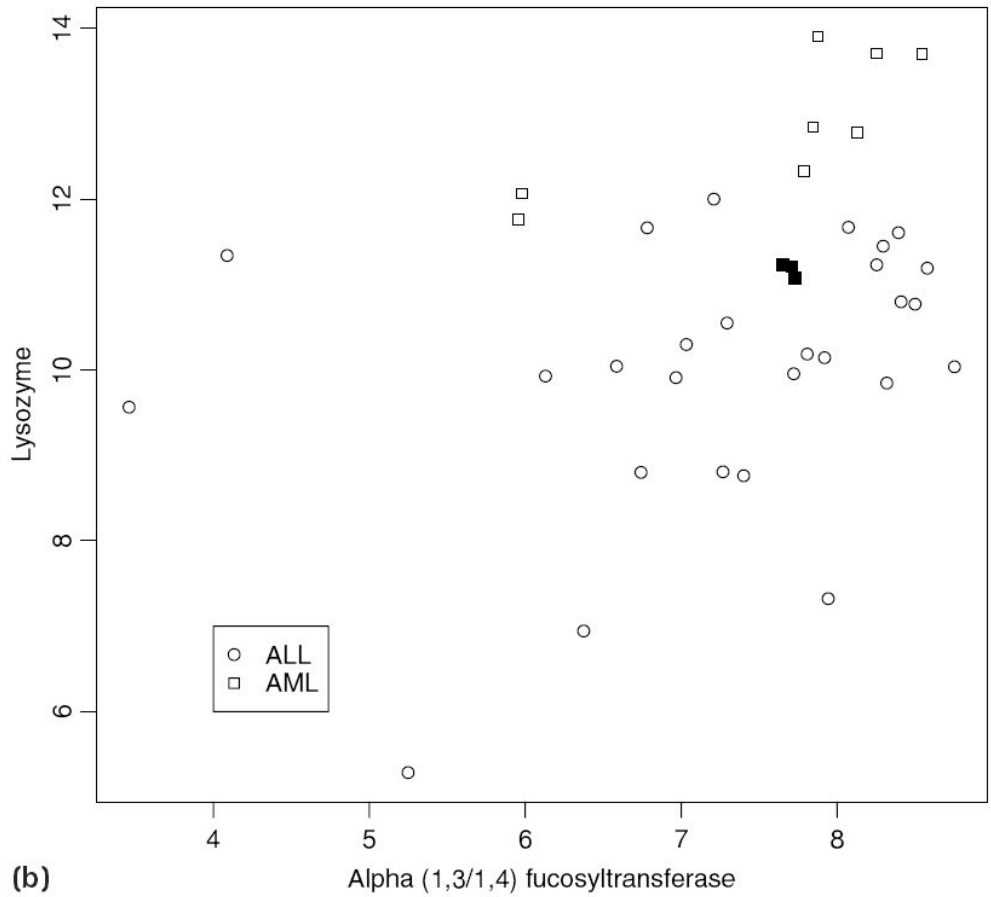
1. Consideriamo le misure di espressione del gene per il campione che stiamo cercando di classificare.
2. Ricerchiamo i campioni conosciuti più vicini tra loro sulla base di una appropriata misura della distanza (tipicamente distanza Euclidea; vedere il paragrafo 8.2).
3. La classe del campione è la classe dei campioni più vicini.

Vi sono due parametri da considerare quando si usi un algoritmo KNN. Il primo parametro è *k*: questo è il numero di campioni più vicini da ricercare; tipicamente viene usato un k = 3, in modo tale che noi si cerchino i tre campioni più vicini, ma alcuni usano k = 5 o k = 1. Il secondo parametro*, l*, è il più piccolo margine di successo per una determinata decisione da prendere; diversamente, gli individui non saranno classificati.

Così se k = 3 ed l = 3, allora i tre elementi più vicini debbono essere nella stessa classe, nell'ambito di una classificazione da fare: se uno dei tre elementi più vicini è in una classe diversa, non sarà fatta nessuna classificazione. Se k = 3 e l = 1, allora un semplice voto di maggioranza condurrà sempre ad una classificazione.


ALL
AML
DNA G/T mismatch binding protein
Immunoglubulin alpha heavy chain
6
7
8
9
10
7
8
9
10
11
12
13

(a)


ALL
AML
Lysozyme
Alpha (1,3/1,4) fucosyltransferase
6
8
10
12
14
4
5
6
7
8

(b)

FIGURA 9.2: Algoritmo k-nearest neighbours. (**a)** l'algoritmo KNN applicato al data set 9A, con k = 3 ed l = 3. Vi sono tre elementi piú vicini dei campioni ALL evidenziati, che sono tutti campioni ALL; pertanto, il campione AML evidenziato ha due elementi piú vicini che sono anche essi campioni AML, ma un elemento che è un campione ALL. Con l = 3, il campione dovrebbe essere non classificato, poiché c'è incertezza sulla classe a cui esso appartiene. Con l = 1 un solo dissenziente sarebbe permesso, e perciò questo campione dovrebbe essere classificato come AML. **(b)** Il problema principale con l'algoritmo KNN è che, in questo caso, i dati non sono molto ben separati. Ci sono tre campioni AML molto vicini e raggruppati insieme in una area in cui è difficoltoso discriminare tra due gruppi. Se noi classificassimo un nuovo campione, e se esso fosse vicino al cluster dei campioni AML, sarebbe classificato come AML, quando noi preferiremmo che esso non fosse classificato.

Reti Neurali[3]

La reti neurali sono un metodo per separare lo spazio in classi in un modo che può includere la separazione non-lineare. Le reti neurali sono basate sul modello di lavoro del cervello: la rete neurale è organizzata in una serie di nodi (che simulano i neuroni), che hanno input ed output (Figura 9.5A). L'uscita dei nodi dipende dall'ingresso nei nodi; differenti ingressi hanno tutti una importanza relativa, che é determinata da un set di parametri denominati *weights* (pesi).

La rete neurale ha la capacità di apprendere aggiustando i pesi. Essa è istruita fornendole esempi di campioni da classificare; la rete aggiusta i pesi dell'input dei nodi così che essa produca una uscita corretta. La rete è istruita fino a che non mostri ulteriori miglioramenti nel predire le classi del set di dati utilizzati per l'istruzione.

Vi sono due passi per usare una rete neurale nella predizione della classe di un individuo:

1. Istruire la rete neurale usando campioni con classe di appartenenza conosciuta.
2. Applicare la rete neurale ad un nuovo individuo per determinarne la classe.

Le reti neurali hanno il vantaggio fondamentale di essere capaci a discriminare dati non linearmente separabili (Figura 9.5b) e possono essere estese naturalmente all'analisi con piú di due classi. Per queste ragioni, le reti neurali stanno diventando strumenti largamente utilizzati in parecchi campi. Di converso, le reti neurali richiedono apprendimento (training) ed ottimizzazione, che rendono questa tecnica piuttosto lenta.

Non esiste una generica architettura per le reti neurali: il numero di nodi nascosti (Figura 9.5a) deve essere identificato empiricamente.

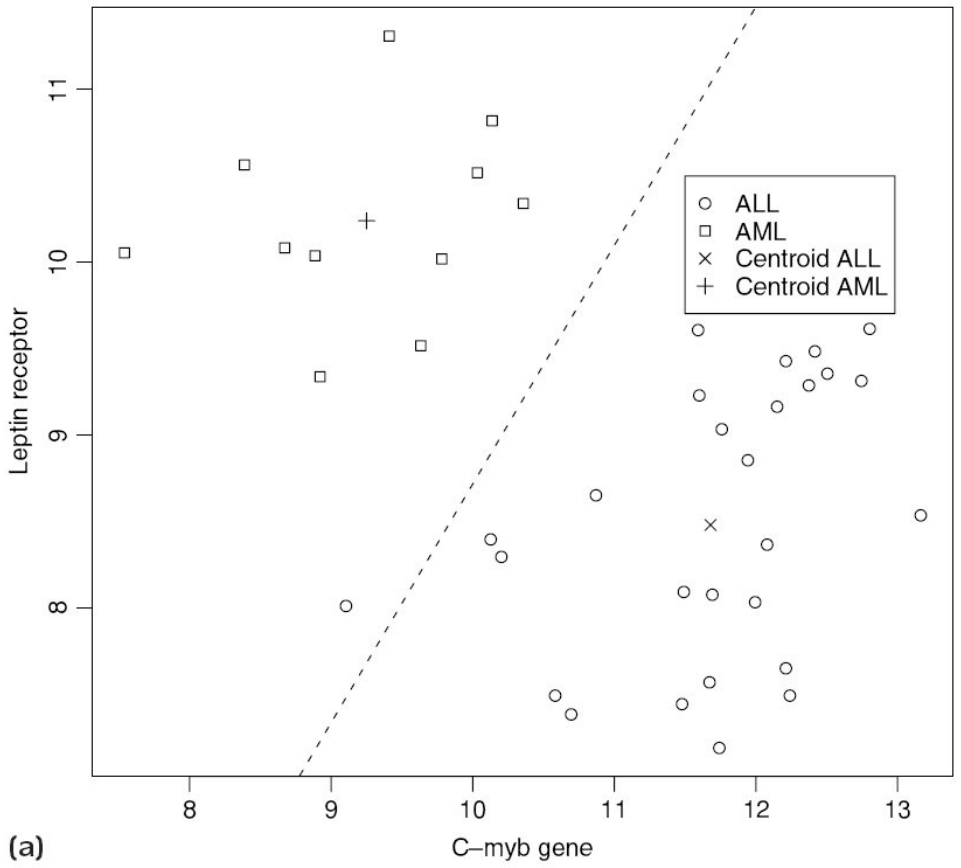
Macchine di Supporto Vettoriali

Le macchine di supporto vettoriale (SVM) costituiscono il metodo più moderno applicato alla classificazione. Le SVM sono simili alle LDA: esse lavorano separando lo spazio in due regioni da una linea retta oppure da un iperpiano con un numero maggiore di dimensioni. L'iperpiano viene scelto in modo tale da minimizzare l'errore di errata classificazione dell'SVM.

La potenza delle SVM è data dal fatto che i dati sono dapprima proiettati in uno spazio con un numero maggiore di dimensioni e quindi separati usando un metodo lineare.

Questo permette una separazione non lineare dei dati. Esiste un numero di modi differenti per proiettare i dati in uno spazio a più alte dimensioni (il metodo scelto viene denominato *funzione kernel* (funzione nucleo). Non andiamo nel dettaglio della matematica in questa sede, ma rimandiamo il lettore alle referenze alla fine del capitolo.

[3] (ndt) Una rete neurale, in sostanza, trova un algoritmo di calcolo non in forma analitica, ma in forma numerica. La grande utilità delle reti neurali é la capacità di ricercare algoritmi di una complessità tale da rendere questi, se esistessero in forma analitica, estremamente complessi ad essere risolti.



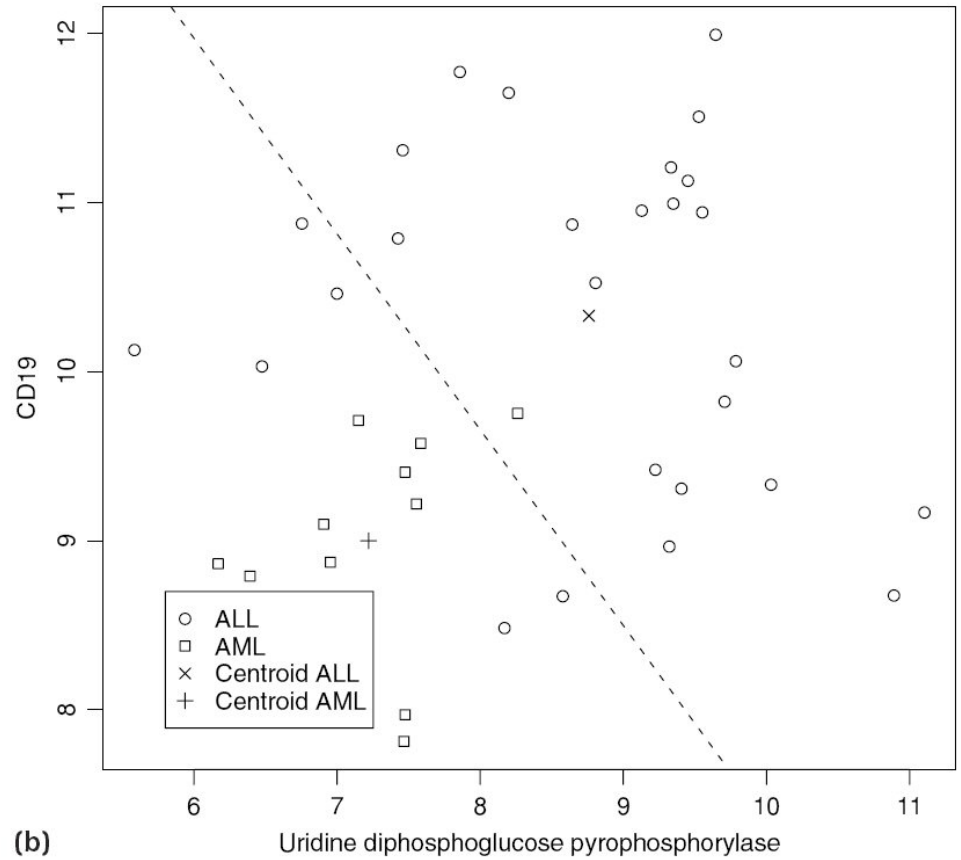


Figura 9.3: Algoritmo del centroide piú vicino. **(a)** calcoliamo i centroidi dei due gruppi; questi sono indicati da una croce per i campioni ALL e con un segno + per i campioni AML. I nuovi campioni sono classificati in base al centroide più vicino. Le due classi sono separate da una bisettrice perpendicolare alla

linea congiungente i due centroidi. In questo caso tutti, tranne uno, dei campioni utilizzati per il training sono classificati correttamente da questo metodo. (**b)** Gli algoritmi basati sui centroidi possono sbagliare con i dati non linearmente separabili. Qui, i geni uridina difosfoglucosio pirofosforilasi e CD19 separano i dati in due regioni, ma i campioni ALL si "*ripiegano*" intorno ai campioni AML. A causa di ciò, molti dei campioni ALL utilizzati per il training sono più vicini ai centroidi AML, e sarebbero classificati erroneamente. Gli algoritmi basati sui centroidi più vicini hanno problemi con i dati separabili non linearmente e dovrebbero essere evitati a meno che non sia chiaro che la separazione delle classi sia lineare.

Vi sono tre passi nel'applicazione della macchina di supporto vettoriale:

1. Proiettare i dati da classi conosciute all'interno di uno adatto spazio a molte dimensioni.
2. Identificare l'iperpiano che separa le due classi.
3. La classe del nuovo individuo è determinata dal lato dell'iperpiano su cui i campioni si allineano.

Le SVM possono discriminare regioni non lineari dello spazio e sono istruite più velocemente rispetto alle reti neurali. Di converso, esse non si estendono naturalmente a più di due classi, e non vi è nessuna funzione kernel naturale da usare: l'operatore deve ottimizzare tutto ciò empiricamente.

9.3 Validazione

Vi è un problema basilare nei metodi di classificazione che abbiamo appena visti: Dato un algoritmo di classificazione, che è stato istruito su due o più gruppi di individui con classi di appartenenza note, come facciamo a sapere che questo metodo sia genericamente applicabile ai nuovi individui? È possibile che gli individui che noi abbiamo usato per istruire l'algoritmo siano in qualche modo non rappresentativi del gruppo a cui essi appartengono e, a causa di ciò, l'algoritmo può fallire nel classificare i susseguenti individui correttamente.

Vi sono due metodi che sono correntemente usati per risolvere questo problema: uso dei **training e test set** e la **cross-validazione (validazione incrociata)**. Se si è in procinto di sviluppare una analisi di classificazione, raccomandiamo l'uso di uno o di entrambi questi metodi di validazione. Training and test sets sono molto efficaci, ma meno potenti con piccoli data set. La Cross-validazione puó essere usata con data set più piccoli ed é tipicamente usata come parte della fase di training, quale aiuto nel'ottimizzare i parametri dell'algoritmo.

Training e Test Set

Questo è il metodo più largamente usato ed accettato per la validazione dei risultati in un algoritmo di classificazione. Approssimativamente due terzi dei dati sono usati per istruire l'algoritmo: l'algoritmo è ottimizzato per classificare il *training data set* nel modo migliore possibile. Dopo il training, l'algoritmo è testato con il rimanente terzo di dati per fornire una verifica indipendente ed una quantificazione di quanto l'algoritmo abbia avuto successo.

Esempio 9.6: Training e Test Set dei dati dal data set 9A

Nel data set 9A, gli autori ebbero a disposizione 62 pazienti per i loro studi, 41 affetti da ALL e 21 affetti da AML. Gli autori scelsero di usare un set di training di 38 pazienti,

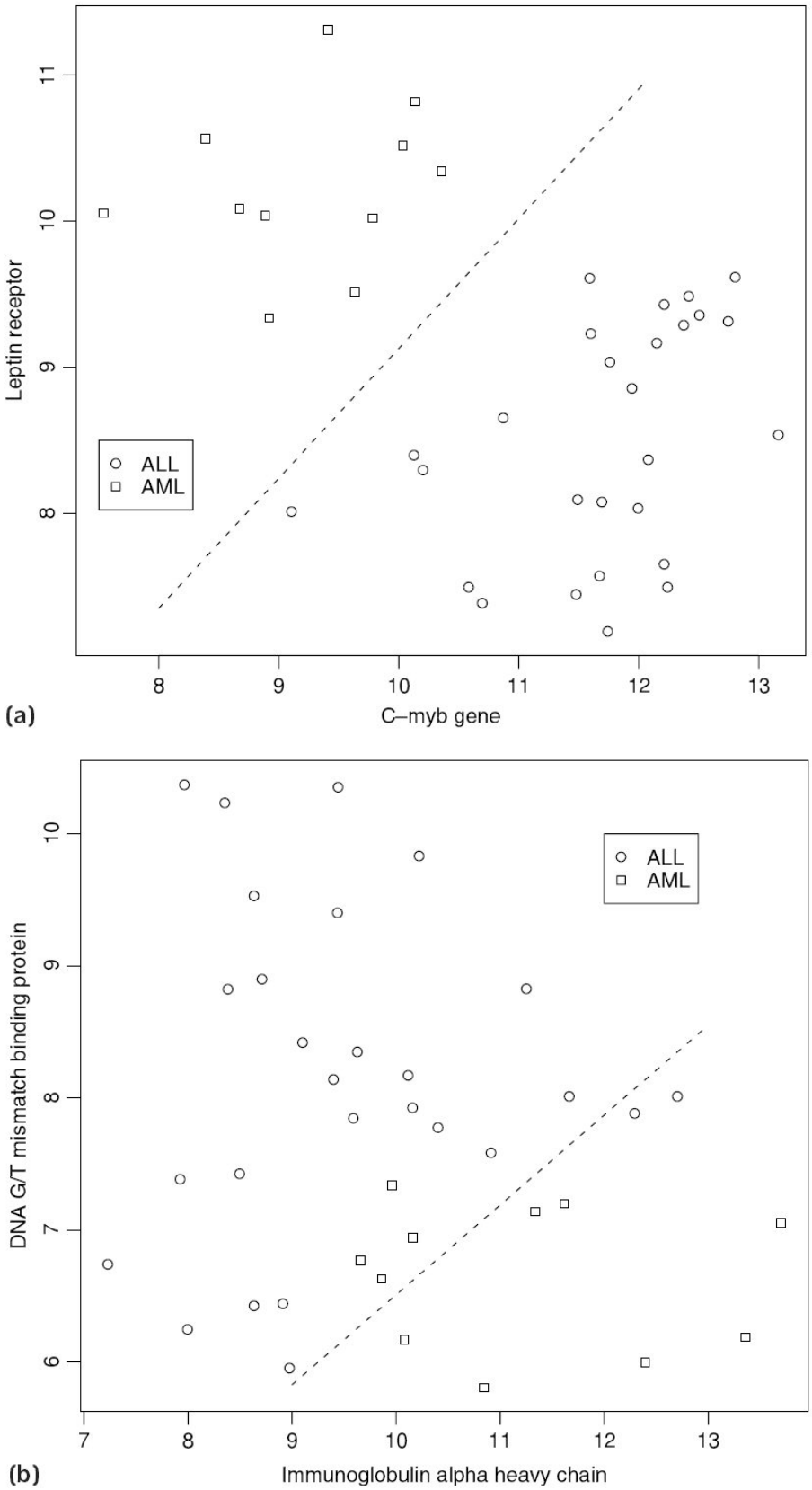


Figura 9.4: Analisi del discriminante lineare (LDA). **(a)** LDA trova la linea retta tra due gruppi che meglio li separa. Quando i gruppi sono linearmente separabili, LDA lavora meglio dell'algoritmo del centroide più vicino, poiché esso tiene conto la variabilità all'interno e tra i gruppi in considerazione (c.f. Figura 9.3a). **(b)** LDA funziona male con i dati separabili non linearmente. Qui i geni sono separati dai dati, ma nessuna linea retta che separi le classi può essere tracciata. Se avessimo usato LDA sui dati originali, parecchi esempi in entrambe le classi sarebbero stati mal classificati. LDA dovrebbe essere evitato, a meno che non sia chiaro che la separazione tra le classi sia lineare.

appena superiore al 60% del data set - 27 pazienti ALL ed 11 pazienti AML. I restanti 24 pazienti (14 ALL e 10 AML) furono usati dagli autori come test set per provare gli algoritmi che essi svilupparono. Il successo di qualsiasi algoritmo puó essere descritto dal numero delle classificazioni corrette nel test set.

Cross-Validazione

L'Alternativa all'uso del *training* e del *test set*, è la *cross-validazione* per misurare il successo di un algoritmo. La cross-validazione ha un parametro (*fold)* associato con esso che determina come l'algoritmo sia stato implementato.

Cross-validazione K-fold divide i dati randomicamente (casualmente) in *k* parti uguali (o quasi uguali). L'algoritmo viene quindi fatto girare *k* volte, usando *k*-1 parti per il training, e la parte restante come test set. Ogni volta che l'algoritmo viene fatto girare, viene usato un diverso test set, così che dopo k iterazioni dell'algoritmo, tutti i dati sono usati come un test set. Il successo dell'algoritmo è la somma delle classificazioni corrette in ciascuna delle iterazioni *runs*.

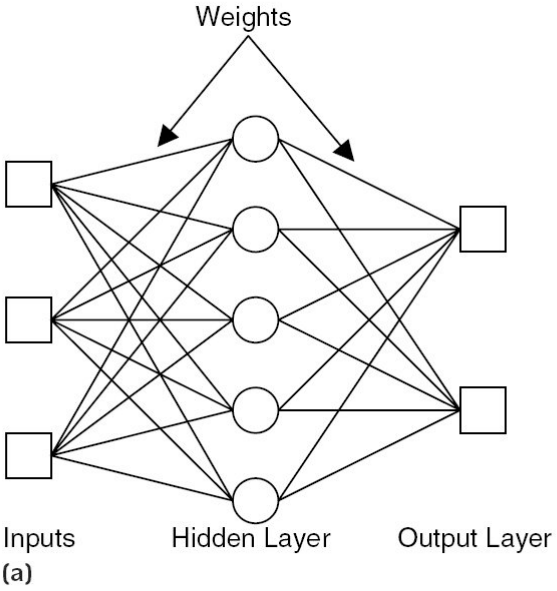
Esempio 9.7: Cross-Validation di un algoritmo di classificazione

Il data set 9A ha 62 pazienti, 41 affetti da ALL e 21 da AML. Una classificazione 3-fold dovrebbe dividere i dati in tre gruppi:

- Gruppo A: 14 pazienti ALL e 7 pazienti AML
- Gruppo B: 14 differenti pazienti ALL e 7 differenti pazienti AML
- Gruppo C: i restanti 13 pazienti ALL ed i restanti 7 pazienti AML

La cross-validation viene fatta girare in tre passi:

1. I gruppi A e B sono usati per il training ed il Gruppo C è usato per il testing.
2. I gruppi A e C sono usati per il training ed il Gruppo B è usato per il testing.
3. I gruppi B e C sono usati per il training ed il Gruppo A è usato per il testing.



(a)

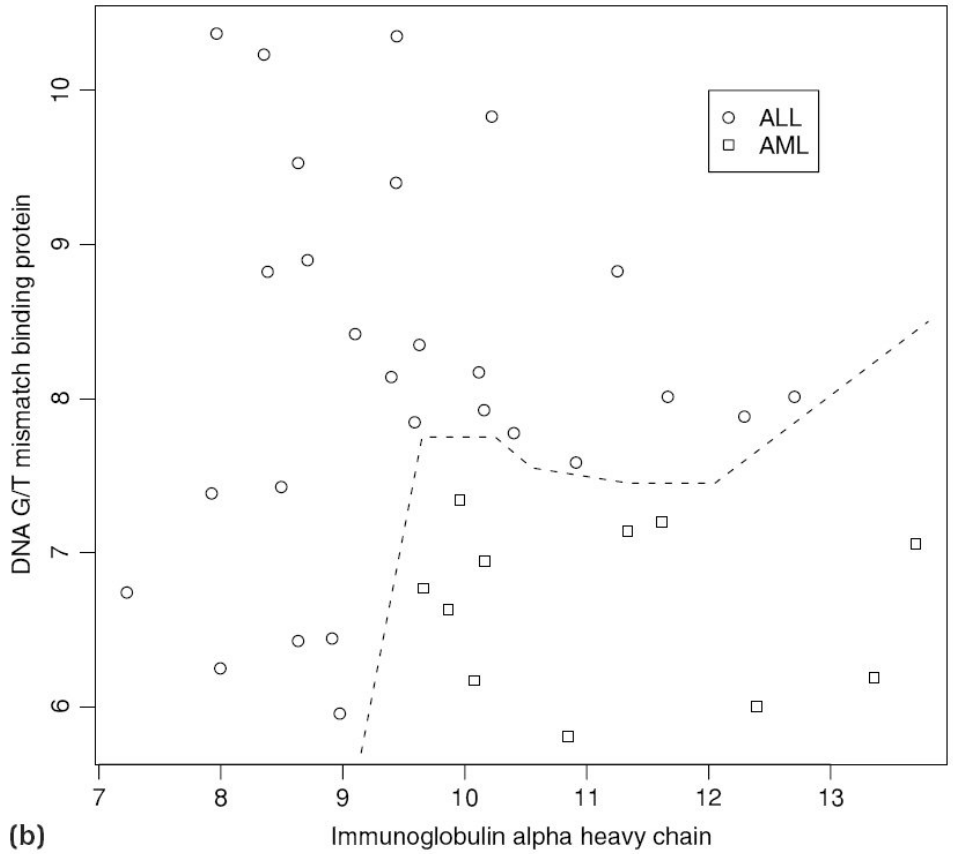


Figura 9.5: Reti Neurali. (a) Una rete neurale è costituita da una serie di nodi ordinati che sono modellati come i neuroni del cervello. Ciascuno degli input é costituito dai geni (o i componenti principali), e si connettono ai nodi dello strato nascosto. Ciascun nodo nello strato nascosto riceve quindi inputs da tutti gli input: ciascun input in ciascun nodo è pesato, e il nodo risponde agli inputs in accordo alla somma dei pesi. I nodi dello strato nascosto sono in condizione di commutare (to fire) oppure di non commutare (not to fire), a seconda che la somma superi, oppure no, una soglia. I nodi dello strato nascosto inviano l'uscita allo strato di output. Questi nodi si comportano esattamente in modo uguale ai nodi dello strato nascosto che abbiamo visto, e commuteranno in accordo alla somma pesata dei loro inputs provenienti dallo strato nascosto. Il numero di classi che la rete neurale può discriminare dipende dal numero dei nodi di uscita: in questo caso, la rete neurale ha due nodi, e quindi può discriminare quattro classi (ndt.: la notazione è booleana, e cioè 00, 01, 10, 11). Il numero di nodi nello strato nascosto determina il campo di azione entro il quale una rete neurale può separare le classi non lineari. La rete neurale *è trained* (*istruita*) mostrando ad essa gli esempi le cui uscite siano note a priori. I *pesi* su tutte le connessioni – sia dall'input fino allo strato nascosto, sia dallo strato nascosto fino all'output- sono aggiustati in modo tale che la rete neurale fornisca l'uscita corretta per ciascuno degli esempi conosciuti. La rete è quindi usata per classificare i campioni la cui appartenenza ai gruppi è sconosciuta. **(b)** La separazione dei pazienti ALL ed AML dal data set 9A usando una rete neurale. La rete neurale è stata in grado di distinguere due gruppi che non erano linearmente separabili. La linea è il confine approssimato tra le regioni di classificazione della rete neurale. Per questo esempio, noi usiamo una rete con soltanto sei nodi nascosti ed abbiamo bisogno soltanto di un nodo di uscita (vi sono soltanto due classi; ndt.: le due classi sono enumerate con la notazione booleana, e cioè 0, 1). Il numero di nodi nascosti necessita di essere determinato empiricamente; questo è lo svantaggio maggiore delle reti neurali.

Lascia-Uno-Fuori Cross-Validation

Un caso speciale particolarmente interessante della Cross-Validation è denominata *n*-fold cross-validation oppure, frequentemente, Lascia-Uno-Fuori cross-validation. In questo metodo, tutti i campioni ad eccezione di uno, vengono usati per creare un classificatore, e l'algoritmo è testato sul campione lasciato fuori. Questo procedimento è ripetuto lasciando fuori, di volta in volta, ciascun campione, ed il numero (o proporzioni) di campioni

correttamente classificati è riportato come un successo dell'algoritmo. La cross-validation è particolarmente utile durante la fase di training dell'algoritmo di classificazione, in cui i parametri possono necessitare di affinamento così che l'algoritmo sia in accordo con il set di dati utilizzati per il training. Di converso, vi è uno svantaggio insito nella cross-validation, che consiste nel fatto che i risultati generati non sono indipendenti e quindi non sono cosí affidabili come quelli che si hanno quando si usi un vero sistema di *training* e *test set*.

9.4 Riduzione della Dimensionalità

Un esperimento con microarray genera dati a dimensionalità molto alta; per esempio, il data set 9A ha 6187 geni. I moderni arrays possono contenere fino a 30000 geni, e ciascun campione ha una misura per ciascun gene. L'esempio che abbiamo visto nel paragrafo 9.2 ha considerato soltanto due geni alla volta, - in uno spazio di misura bidimensionale - ma gli esperimenti con microarray da cui provengono questi esempi hanno misure in uno spazio dimensionale molto più alto: parecchie migliaia di dimensioni in ciascuno dei due data set che abbiamo descritto. Uno dei problemi implicati nel costruire un buon classificatore è quello della riduzione della dimensionalità di dati: invece di considerare tutte le 6000 misure di espressione dei geni di un campione, noi consideriamo un piccolo numero di misure. Nell'esempio che abbiamo mostrato prima, abbiamo considerato appena due geni; in realtà, possiamo anche desiderare di usare piú geni, forse da 5 a 20 geni. Vi è un certo numero di ragioni che ci inducono a ridurre la dimensionalità del sistema:

- **Rimozione del rumore e della informazione irrilevante.** Parecchi geni non contengono informazione utile per determinare le differenze tra i campioni. Questi geni non dovrebbero essere usati per la classificazione; invero, qualche volta essi possono contenere perfino rumore che puó portarci ad una scorretta classificazione.
- **velocità di training dei metodi**. Un certo numero di metodi che abbiamo descritto, come per esempio le reti neurali, funzionano meglio con meno informazione in ingresso. È necessario, dunque, ridurre la dimensionalitá dei dati prima che si possa usare questi metodi con profitto.
- **Informazione identica.** Alcuni geni sono altamente correlati e contengono, dunque, la stessa informazione. L'inclusione di tutti questi geni può essere causa della inaffidabilità di qualche metodo.
- **Molteplicità**. Quando stiamo considerando alcune migliaia di geni in parallelo, puó succedere che qualcuno di questi geni sia differenzialmente espresso tra differenti campioni, ma di fatto queste differenze possono essere dovute a variazioni random.
- **Tool diagnostici**. Di solito, lo scopo è quello di produrre un tool diagnostico o prognostico per le malattie o i trattamenti che si stanno studiando. Mentre potrebbe essere utile usare il microarray come un tool generale, in molti casi sarà piú economico e più efficiente produrre un tool di utilizzo più focalizzato, come la PCR quantitativa, che usa pochi, e piú rilevanti, geni.
- **Generazione di ipotesi.** Una classificazione basata su un piccolo numero di geni può essere la base di una ipotesi scientifica circa il ruolo dei geni rilevanti in diverse malattie o trattamenti che si stanno studiando. Per far ciò, è necessario trovare tali geni.

TABLE 9.2: Advantages and Disadvantages of Four Dimensionality Reduction Methods

Principal Component Analysis	Individual Gene Selection	Pairwise Gene Selection	Genetic Algorithms
✓ Quick and easy to use × Does not provide a subset of classifying genes × Principal components may not separate the classes	✓ Quick and easy to implement ✓ Generates a subset of classifying genes × Best individual genes may not make the best classifiers × Need to combine selected genes to produce classifier	✓ Can generally find good classifiers ✓ Generates a subset of classifying genes × Need to combine selected genes to produce classifier × Slower than individual gene selection	✓ Finds the best classifiers ✓ Generates a subset of classifying genes × Slowest algorithm × Need programming skills to implement

La selezione di un subset appropriato di geni è un difficile problema ed è soggetto a continue ricerche. Nella teoria della **computer science**, un problema è classificato come *hard (notevole)* se il numero di passi per valutare la soluzione incrementa esponenzialmente con la dimensione del problema. In questo caso, il numero di possibili subset (sottoinsiemi) di N geni è 2^N, così che la valutazione di tutti i possibili gruppi di geni incrementa esponenzialmente con il numero di geni che si stanno studiando.

I quattro metodi che descriviamo sono tutti comunemente usati e di solito si ritrovano in letteratura. Tutti i metodi hanno vantaggi e svantaggi. I metodi che descriviamo sono:

- Analisi delle componenti principali
- Selezione individuale del gene
- Selezione dei geni presi a coppia a coppia
- Algoritmi genetici

Per illustrare questi metodi li applicheremo al data set 9A, usando in particolare quei metodi di classificazione descritti nel paragrafo 9.2 ed, inoltre, useremo il *training* ed il *test set* dell'esempio 9.6. Nella tabella 9.2, riassumeremo i vantaggi e gli svantaggi di ciascuno di questi metodi.

Analisi delle Componenti Principali

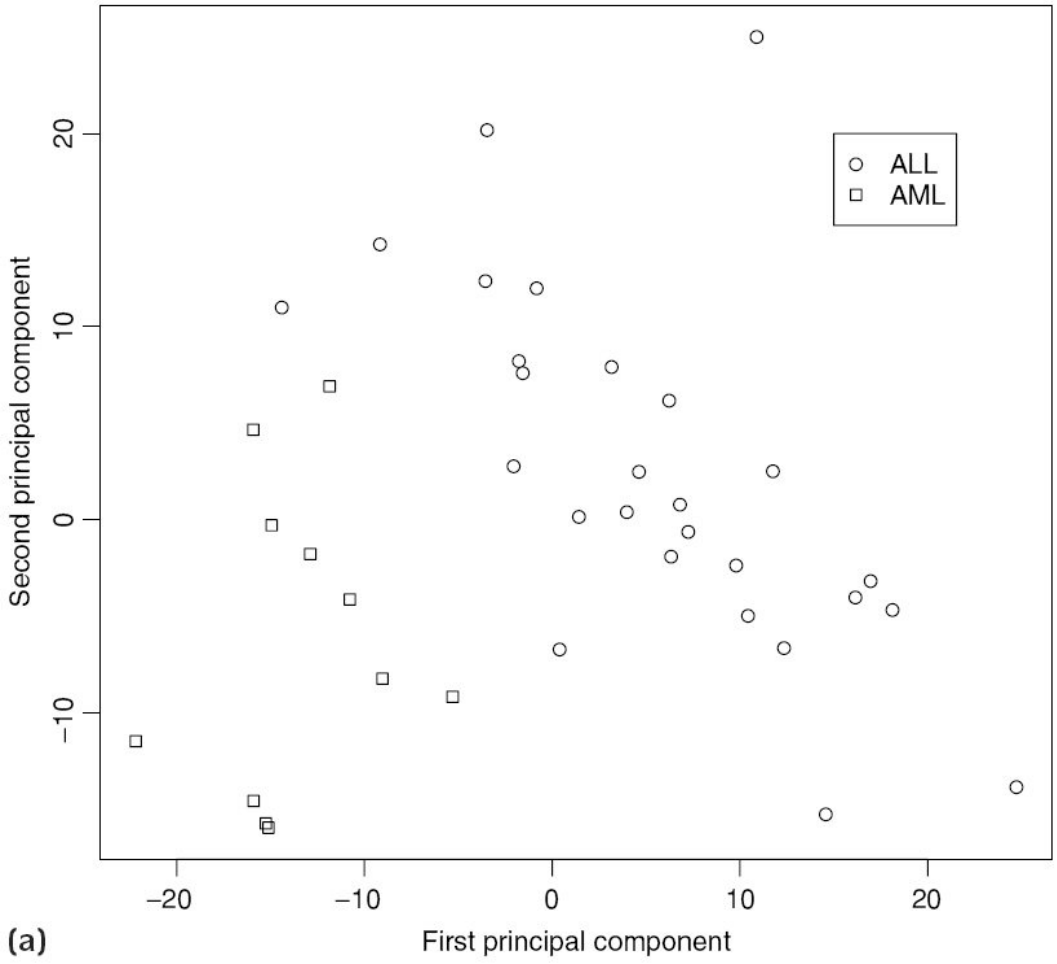
Il primo metodo che descriviamo non trova - nel vero senso del termine - un sottoinsieme di geni rilevanti, ma usa l'analisi delle componenti principali (PCA) per ridurre la dimensionalità dei dati. La PCA è descritta estesamente nel paragrafo 8.3; è immediato usarla poiché essa è implementata in tutti i pacchetti di analisi dell'espressione genica, oltre che in molti pacchetti di analisi avanzata dei dati come, ad esempio, R o Mathlab.

Esempio 9.8: PCA applicato al data set 9A

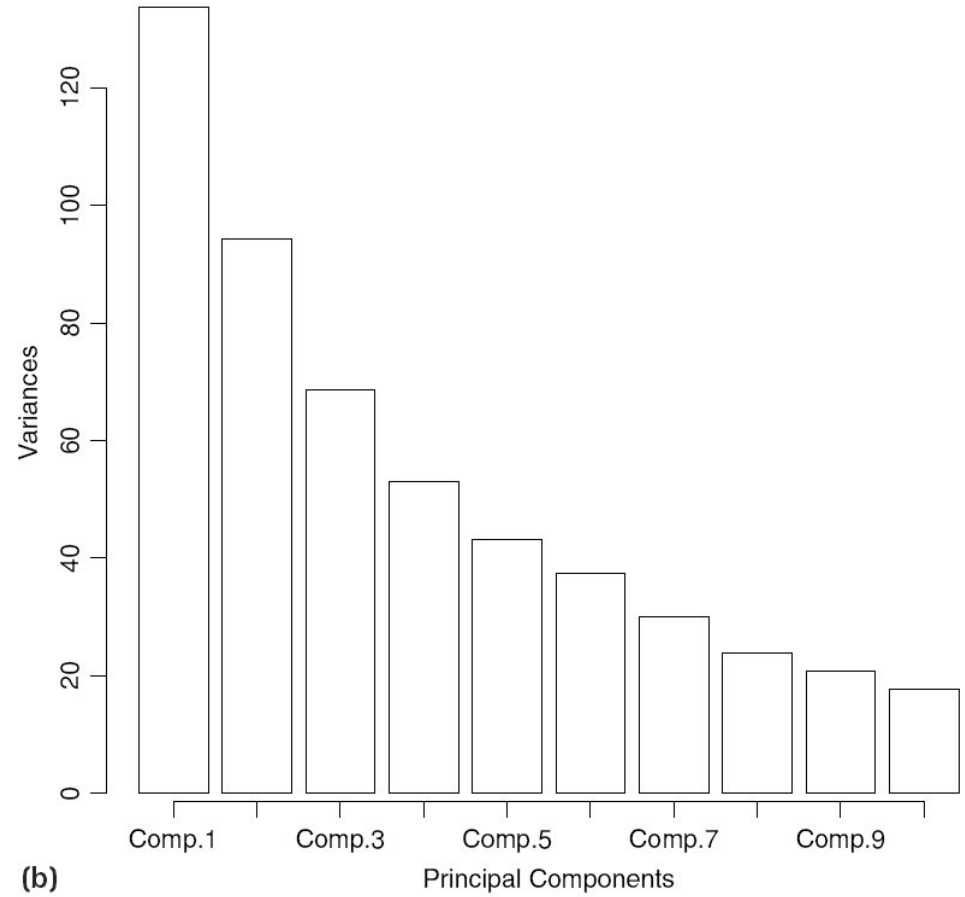
Quando la PCA è applicata al data set 9A, molti dei metodi di classificazione (ad eccezione di KNN) sono in grado di separare i dati (Figura 9.6; Tavola 9.3). Di converso benché questo sia vero per questo particolare data set, non è necessariamente vero per altri data set.

Vi sono parecchi fattori che devono essere considerati quando si usi la PCA per la classificazione:

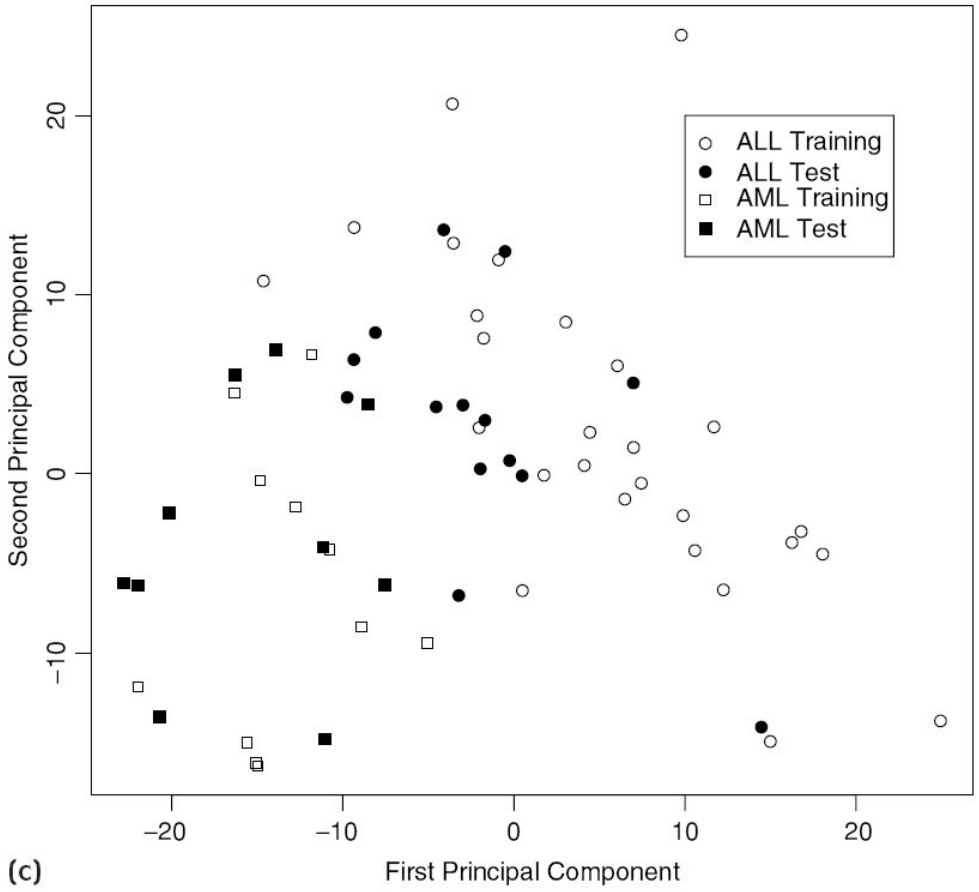
- Il metodo PCA richiede tuttora le misure di un elevato numero di geni. Se lo scopo è quello di avere un piccolo numero di geni che possa essere misurato per un futuro uso diagnostico, allora la PCA non è adatta.
- La PCA trova gli assi che catturano la variabilità dei dati in generale. Se le classi sono separate per mezzo delle componenti principali, allora la PCA è un metodo eccellente; se le componenti principali non separano le classi, la PCA deve essere abbandonata e deve essere usato un metodo differente.
- La PCA è basata sulla combinazione lineare dei geni; se sono necessarie combinazioni non lineari dei geni per separare i dati, allora PCA non funziona.



(a)


Variances
0
20
40
60
80
100
120
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9
Principal Components

(b)


ALL Training
ALL Test
AML Training
AML Test
Second Principal Component
−10
0
10
20
−20
−10
0
10
20
First Principal Component

(c)

Figura 9.6: Analisi del componente principale (PCA). PCA è un metodo popolare per ridurre la dimensionalità della matrice di espressione dei geni applicando un algoritmo di classificazione. In questo esempio, è stato applicato al data set 9A. **(a)** I due gruppi di pazienti (ALL ed AML) disegnati nei primi due componenti principali. In questo caso, i dati sono separabili dai componenti principali. Questo non è necessariamente sempre vero con i dati dei microarray e per questa ragione le componenti principali non sono necessariamente il miglior metodo per la riduzione della dimensionalità. **(b)** le primi 10 componenti principali contribuiscono tutti ad una generale variabilità dei dati. Dovrebbe avere un senso includere tutte queste componenti nell'analisi di classificazione. Altri data set (e.g., data set 8A) hanno poche componenti principali che contribuiscono ad una variabilitá generale (Tabella 8.3), cosí in questo caso uno dovrebbe usare un piccolo numero di componenti principali nell'analisi di classificazione. **(c)** dati provenienti da un test set indipendente sono mappati nelle prime due componenti principali del training set. In generale, la separazione è buona; vi è un piccolo numero di campioni nella regione di confine che sarebbe classificata in modo scorretto dalle prime due componenti principali.

TABLE 9.3

Predictions of different algorithms applied to data set 9A using the 10 most significant principal components and using 38 patients in a training set and 24 patients in a test set. The number of samples in each of the classes is given in parentheses. With most algorithms, PCA worked very well. This is not always the case.

Method	Training ALL (27)	Training AML (11)	Test ALL (14)	Test AML (10)
Nearest centroid	26	11	14	10
KNN ($k = 3, l = 3$)	26	11	12	3
LDA	26	11	13	10
Neural network (10 HNs)	27	11	14	8
Neural network (20 HNs)	27	11	14	10

Selezione Individuale del Gene

Il metodo più semplice che realmente sceglie i geni consiste nell'assegnare un rango ai geni che sono individualmente discriminati nel migliore dei modi tra le due classi e poi usare i migliori geni come classificatori. Una buona misura di discriminazione che è usata comunemente è la il t-statistic (paragrafo 7.3). Questo cattura la differenza tra la media delle classi come rapporto della deviazione standard dei due gruppi.

Esempio 9.9: Geni individuali con rango più alto per il data set 9A

I geni che possono essere discriminati tra il pazienti ALL ed AML del data set 9A

Hanno tutti dei *p-values* molto buoni associati con i loro t-*statistic* (Tabella 9.4). Di converso, perfino il miglior gene non separa le classi molto bene in confronto con la separazione che puó essere ottenuta da due o piú geni (Figura 9.7). Due dei geni in questa tabella, C-myb ed il recettore della leptina, formano un buon classificatore come coppia di geni (Tabella 9.5). Di converso, altri geni che formano buoni classificatori di coppia non sono necessariamente buoni classificatori individualmente.

TABLE 9.4: Top 10 Genes That Individually Classify the AML and ALL Patients

Gene	*t*-test *p*-value
CD33 antigen	1.9E-09
C-myb gene	6.32E-09
Leptin receptor	8.93E-08
Cathepsin D	1.99E-07
Transcription factor 3	2.02E-07
Connective tissue activation peptide III	3.48E-07
Myosin light chain	4.19E-07
Granulin	4.31E-07
Retinoblastoma binding protein P48	5.32E-07
NADPH-flavin reductase	6.85E-07

Nota: I dieci geni principali in accordo al p-value ed al t-test applicati ai due gruppi su base gene per gene (Paragrafo 7.2). I geni sono stati prefiltrati per includere soltanto quei geni che sono stati espressi in tutti i pazienti. Si noti che in questo caso particolare, i geni C-myb e recettore della leptina sono buoni predittori delle due classi (Tabella 9.6). Di converso, molti degli altri geni che sono buoni predittori nelle coppie non sono necessariamente buoni come predittori di classi come singoli.

TABLE 9.5

Predictions of different algorithms on data set 9A, using the training and test sets of Example 9.6 and using the 10 best genes as determined by single-gene classification (Table 9.4). In this case, the training data are linearly separable using the 10 best genes; we have already seen this in Figure 9.1b, where the genes C-myb and leptin receptor separate the classes. Using these genes, all methods have performed comparably, but no methods have been able to predict all of the test AML cases correctly. It would appear likely that the test data are not separable with these 10 genes.

Method	Training ALL (27)	Training AML (11)	Test ALL (14)	Test AML (10)
Nearest centroid	27	11	14	8
KNN ($k = 3, l = 3$)	27	11	14	7
LDA	27	11	14	8
Neural network (10 hidden nodes)	27	11	14	5
Neural network (20 hidden nodes)	27	11	14	7

I metodi di classificazione descritti nel Paragrafo 9.2 possono essere usati con i primi geni discriminanti (Tabella 9.4) per costruire un modello di classificazione per il data set 9A , sia direttamente (Tabella 9.5), sia usando un algoritmo di voto (vedere la discussione che segue). Benché questo metodo abbia funzionato bene con il training set, esso non ha funzionato così bene con i dati di test, con solo tra 5 ed 8 pazienti del gruppo AML classificati correttamente. È come se i dati di training non fossero separabili usando questi 10 geni.

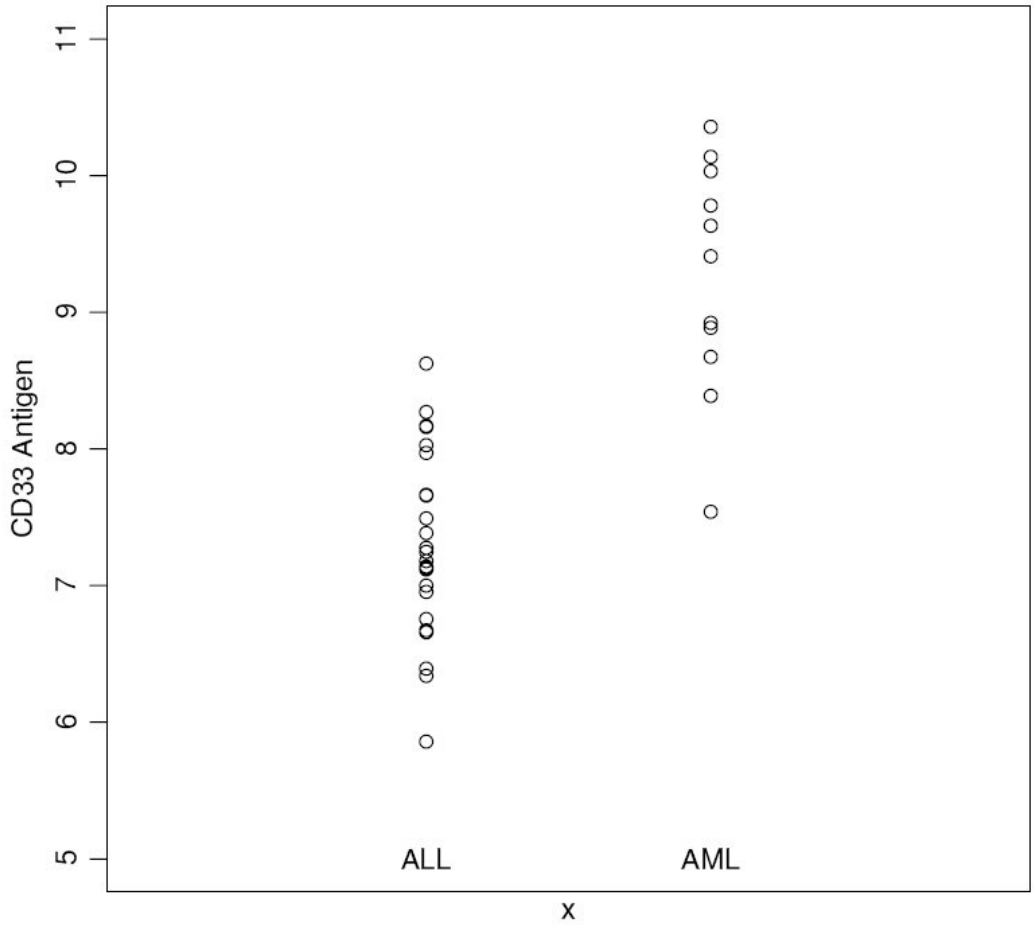


Figura 9.7: Classificazione di geni individuali. Il gene che meglio distingue tra i pazienti ALL ed AML come da misure di statistiche-t è l'antigene CD33 (Tabella 9.4). Esso non separa le classi molto bene: vi è una sostanziale sovrapposizione dei valori di espressione del gene. Metodi per la selezione del gene basati sulla scelta dei migliori geni individuali non sono generalmente molto buoni.

In generale, la selezione dei geni individuali è il metodo più debole per selezionare i geni da usare per la discriminazione. La ragione di tutto ciò è che i geni che sono buoni classificatori individuali possono non lavorare bene insieme per classificare campioni. Inversamente, i geni che non lavorano bene insieme per classificare i campioni possono non essere buoni individualmente. Per questa ragione, si raccomanda di usare un metodo che consideri le coppie di geni, oppure i gruppi di geni, per classificare i dati.

Selezione dei geni coppia a coppia

Questo è un metodo più sofisticato, e non fa affidamento soltanto su un singolo gene; esso considera le coppie di geni che sono le migliori in grado di discriminare i campioni usando il metodo scelto, e quindi combina insieme questi geni per generare un predittore generale. Per esempio, con l'algoritmo KNN, applicheremmo l'algoritmo a tutte le coppie di geni, e quindi selezioneremmo -diciamo- le cinque migliori coppie di geni per un predittore a 10-geni. Un approccio simile potrebbe essere usato per qualunque altro metodo.

Esempio 9.10: Selezione dei Geni Coppia a coppia usando KNN

Usando il data set 9A, l'algoritmo KNN è stato applicato ai 500 geni più variabili per identificare le dieci coppie di geni che ciascuna classifica correttamente con 37 pazienti su 38 in una cross-validazione del training set (Tabella 9.6). Questi corrispondono a 14 geni unici, che possono essere usati per costruire un classificatore.

Se KNN è applicato usando 14 geni, allora ogni campione nel training set è classificato correttamente. Di converso, l'algoritmo funzioni meno bene sul test set, producendo un risultato corretto solo 15 volte su 24 (Tabella 9.7).

La selezione dei geni in coppia può essere usata per ridurre la dimensionalità di tutti i metodi descritti nel paragrafo 9.2. In generale, la selezione in coppie di geni funziona meglio rispetto alla selezione di un singolo gene (Tabella 9.7). Di converso, essa è più lenta della selezione di un singolo gene poiché il numero delle coppie di geni è la metà del quadrato del numero di geni.

TABLE 9.6

Pairwise gene selection has been applied to the training set of data set 9A using the KNN algorithm. There are 10 pairs of genes that correctly predict 37 out of 38 patients in a cross-validation of the training set. There are 14 unique genes in this list, which can be used as genes for classification of these tissues.

Genes		Training ALL (27)	Training AML (11)
Ferritin heavy chain	CD33 antigen	27	10
RLIP76 protein mRNA	CD33 antigen	27	10
Casein kinase 1 delta	CD33 antigen	27	10
DNA-damage-inducible transcript 1	CD33 antigen	26	11
NADPH-flavin reductase	CD33 antigen	27	10
LEPR leptin receptor	C-myb gene	26	11
Cholinergic receptor, nicotinic, alpha polypeptide 7	C-myb gene	26	11
Cholinergic receptor, nicotinic, alpha polypeptide 7	Topoisomerase (DNA) II beta	27	10
Catalase	Cytoplasmic dynein light chain 1	26	11
Nucleoside-diphosphate kinase	Retinoblastoma binding protein P48	27	10

TABLE 9.7

Predictions of different algorithms using the 10 best genes selected by pairwise gene selection on data set 9A. Most of the algorithms performed better with pairwise gene selection than with individual gene selection.

Method	Training ALL (27)	Training AML (11)	Test ALL (14)	Test AML (10)
Nearest centroid	26	11	14	10
KNN ($k = 3, l = 3$)	27	11	11	4
KNN ($k = 3, l = 0$)	27	11	13	8
LDA	26	11	13	10
Neural network (10 HNs)	27	11	14	8
Neural network (20 HNs)	27	11	14	10

Algoritmo di voto

Vi sono due modi di combinare la classificazione dei geni che sono selezionati usando l'algoritmo di un singolo gene oppure l'algoritmo di selezione a coppie di geni. Il primo metodo consiste nell'usare l'algoritmo con tutti i geni insieme. Nell'Esempio 9.10, vi sono 10 coppie di geni selezionate per KNN, con 14 geni tra di essi. Questi 14 geni possono essere usati per costruire un singolo predittore KNN.

Il secondo metodo consiste nel'uso dell'algoritmo di voto. In un algoritmo di voto, ciascun campione è classificato da ciascun gene o coppie di geni selezionate, e la classificazione maggioritaria è usata come classe di quel gene.

Esempio 9.11: Uso dell'algoritmo di voto con coppie KNN

10 coppie di geni sono usate per classificare, di volta in volta, ciascuna i 24 campioni di test. Il primo campione ALL è classificato 8 volte come ALL, una volta non classificato, ed una volta come AML. Tutti e 14 i campioni ALL sono stati correttamente classificati, mentre 6 dei 10 campioni AML sono stati correttamente classificati (Tabella 9.8).

TABLE 9.8

The 14 test samples are classified using voting from the 10 pairs selected by KNN (Table 9.6). In each row, we count the number of times the individual was classified as ALL, unclassified or classified as AML by the 10 pairs. All ALL samples are correctly classified; there is more difficulty with the AML samples – 6 out of 10 are correctly classified.

Sample	ALL	Unclassified	AML
ALL 1	8	1	1
ALL 2	7	3	0
ALL 3	8	0	2
ALL 4	9	0	1
ALL 5	10	0	0
ALL 6	10	0	0
ALL 7	10	0	0
ALL 8	9	1	0
ALL 9	10	0	0
ALL 10	9	1	0
ALL 11	10	0	0
ALL 12	9	1	0
ALL 13	10	0	0
ALL 14	10	0	0
AML 1	0	1	9
AML 2	0	2	8
AML 3	0	0	10
AML 4	4	5	1
AML 5	3	3	4
AML 6	2	3	5
AML 7	6	4	0
AML 8	4	4	2
AML 9	1	2	7
AML 10	7	2	1

Algoritmi Genetici

Il metodo più potente che descriviamo per la scelta dei sottoinsiemi di geni è un metodo denominato *Algoritmo Genetico*. Gli algoritmi genetici sono metodi computazionali atti a risolvere problemi difficili, ed hanno la loro ispirazione alla biologia evoluzionistica.

In biologia, organismi con differenti genotipi hanno differenti fenotipi, che sono più o meno adatti, e che passano alla prole nella successiva generazione. Negli algoritmi genetici, vi è una intera popolazione di soluzioni ad un dato problema; ciascuna soluzione ha un "genotipo", che descrive i parametri della soluzione. L'idoneità della soluzione è l'abilità dell'algoritmo a risolvere il problema. Il maggiore adattamento degli individui sono selezionati in ciascuna generazione per produrre la prole nella successiva generazione.

Così come con la biologia reale, le soluzioni individuali possono produrre prole sia asessuata che sessuata. Con la riproduzione asessuata, la prole è identica ai genitori, con la possibilità di cambiamenti dovuti a mutazioni casuali. Con la riproduzione sessuale. I "genomi" (parametri) di due individui sono combinati insieme per produrre un nuovo individuo. Descriviamo un semplice algoritmo genetico che può essere usato per scegliere un subset di geni per l'analisi della classificazione. Come vedremo, vi sono molti modi in cui un algoritmo genetico può essere implementato, e questi possono essere usati con qualsiasi metodo di classificazione. La implementazione che descriviamo non è necessariamente la migliore: essa è inclusa per scopi dimostrativi, e descriveremo, in aggiunta, alcune possibili modificazioni.

L'Algoritmo

Vi è una popolazione di N individui, ciascuno dei quali ha un genoma consistente in una lista di N geni che saranno usati insieme come classificatori. Vi sono cinque passi:

1. Ciascun individuo parte con una scelta casuale di n geni.
2. Si costruisce una nuova, e più grande popolazione rispetto alla vecchia; questa sarà tipicamente di grandezza 3N, usando tre metodi di riproduzione:
    a. Clonazione. N individui nella nuova popolazione sono uguali ad N individui della vecchia generazione.
    b. Mutazione. N individui sono creati da ciascuno degli N individui della vecchia generazione, che sono identici a ciascuno dei loro genitori, ma con un gene cambiato casualmente.
    c. Ricombinazione. N individui sono creati selezionando in modo casuale due genitori dalla precedente generazione, combinando i loro geni in un singolo pool, e quindi selezionando n geni dal pool combinato.
3. Calcolo della *fitness* di ciascuno dei 3N individui. In questo caso, la *fitness* sarà il numero dei campioni che sono classificati correttamente in una cross-validazione del tipo *lascia uno fuori* del metodo di classificazione applicato al training set usando n geni di quell'individuo.
4. Selezionare i migliori N individui per formare la successiva generazione.
5. Ritornare al passo 2 e continuare fino a quando la popolazione contiene soluzioni sufficientemente buone.

Esempio 9.12: Algoritmo Genetico con KNN

Come esempio, applichiamo questo algoritmo genetico per selezionare un gruppo di 8 geni che classificherà i campioni di ALL ed AML del data set 9A. In questo caso, utilizziamo una grandezza della popolazione di 50, ed abbiamo 8 geni in ciascun classificatore. Nella settima generazione della simulazione I ran, vi era un classificatore che riportò 38 su 38 classificazioni corrette sulla cross-validation del training set (Tabella 9.9). Quando questo classificatore fu applicato al test set, tutti i 14 pazienti ALL furono classificati correttamente, e 7 su 10 pazienti AML furono allo stesso modo classificati correttamente.

Questo algoritmo è solo un esempio di come un algoritmo genetico dovrebbe essere implementato per risolvere questo problema. Vi sono parecchie modifiche che potrebbero essere applicate all'algoritmo, includendo:

- Variazione dell'ampiezza della popolazione;
- Variazione del numero o proporzione della prole creata a ciascuna generazione per mezzo della clonazione, mutazione o ricombinazione;
- Permettere a differenti individui di avere differenti numero di geni, possibilmente includendo l'aumentata fitness per la classificazione usando pochi geni.

Gli algoritmi genetici sono largamente applicabili a molti difficili problemi computazionali. Di converso, vi è anche un costo: gli algoritmi genetici sono notoriamente lenti, poiché essi usano eventi random per generare la generazione successiva.

Riassunto dei punti chiave

Vi sono parecchi metodi per classificare i campioni, inclusi:

- K-nearest neighbours (elementi attorno piú vicini)
- Classificazione con i centroidi
- Analisi del discriminante lineare
- Reti neurali
- Macchine di supporto vettoriale

Con i dati dei microarrays, abbiamo bisogno di ridurre la dimensionalità dei dati per trovare gruppi di geni che siano in grado di separare i dati. Vi sono parecchi metodi per fare ciò, includendo:

- Analisi del Principale Componente
- Selezione del gene individuale
- Selezione di coppie di geni
- Algoritmo genetico

L'Analisi di classificazione dovrebbe essere verificata usando il training ed i test sets e/oppure la cross-validation (validazione incrociata).

***Capitolo* 10**

Progetto di un Esperimento

10.1 Introduzione

La progettazione di esperimenti con microarrays é una delle più importanti aree dell'informatica dei microarrays ed è un argomento di vecchia data nell'ambito della statistica classica. La ragione per fare buoni progetti sperimentali è che essi permettono di ottenere la massima informazione da un esperimento con minimo sforzo - che ha, peraltro, un risvolto in tempo e denaro. L'alternativa ad un buon progetto sperimentale è quella di sviluppare esperimenti con i microarrays che producono dati non utilizzabili. Ci si potrebbe chiedere perché introduciamo questo argomento a questo punto del libro, dopo l'analisi dei dati, piuttosto che all'inizio del libro, insieme al materiale relativo alla progettazione dei microarrays stessi. Vi sono due ragioni per tutto ciò. La prima è che gli argomenti di questo paragrafo usano concetti che sono stati sviluppati nei precedenti capitoli; argomenti notevolmente importanti quali, ad esempio, i *test di ipotesi* ed il *p-values* introdotti al capitolo 7. Ma vi sono anche ragioni più filosofiche perché io abbia scelto di mettere il materiale relativo al progetto di esperimenti dopo il materiale sull'analisi dei dati. Dal mio punto di vista, è assolutamente fondamentale comprendere le questioni scientifiche alle quali si deve provare a dare una risposta, e perfino le ipotesi scientifiche che si cerca di formulare, prima di progettare l'esperimento. A questo punto, pertanto, si dovrebbe essere giunti a possedere una idea chiara della struttura dei dati che si cerca di produrre e del tipo di analisi che si intende impiegare prima del progetto dell'esperimento.

Questo capitolo considera tre aree del progetto sperimentale:

*Paragrafo 10.2***:** Blocco (Blocking), Randomizzazione (dati casuali) ed Acciecamento (dei dati), pongono l'attenzione ai problemi statistici di confusione e polarizzazione (bias), ed ai metodi che vengono usati per risolvere questi problemi.

Paragrafo 10.3: Scelta della Tecnologia di Strutturazione dei Campioni, discute i benefici relativi di Affimetrix e delle piattaforme di campioni a due colori; e la distribuzione dei campioni sull'array in un certo numero di tipi di esperimenti con microarrays.

Paragrafo 10.4: Quanti Replicati? , descrive i metodi statistici per determinare il numero di replicati che sarebbero necessari all'uso di esperimenti con microarrays per ottenere quei dati che possano rivelare gli effetti che stiamo ricercando.

10.2 Blocco, Randomizzazione e Acciecamento

Introduciamo l'argomento con un semplice esempio, di cui descriveremo tre progetti sperimentali: sui primi due sorvoliamo velocemente, mentre il terzo lo esamineremo con una certa profondità.

Esempio 10.1: Risposta tossica al BENZO(A)PiRENE

In un esperimento per investigare gli effetti del benzo(a)pirene, una epatotossina nota, 8 ratti verranno trattati con benzo(a)pirene, ed 8 ratti verranno trattati con la sostanza di controllo. Campioni di fegato verranno preparati da tutti e 16 i ratti ed ibridizzati con 16 arrays. La preparazione dei campioni e la ibridizzazione saranno sviluppate da due ricercatori, Alison e Brian, ed i 16 arrays provengono da due lotti di 8 arrays prodotti in due cicli diversi.

Disegno Sperimentale 1

Alison scelse 8 ratti e li trattò con benzo(a)pirene. Ella preparò i campione di fegato dai ratti e li ibridizzò a 8 arrays provenienti dal primo lotto. Brian prese i rimanenti ratti e li trattò con la sostanza di controllo; egli preparò i campioni e li ibridizzó ad 8 arrays dal secondo lotto.

Disegno Sperimentale 2

Alison scelse 8 ratti, e ne trattò 4 con benzo(a)pirene e 4 con la sostanza di controllo. Ella scelse 4 arrays per ciascuno dei lotti, ed ibridizzó i campioni da due ratti trattati e da due ratti di controllo a ciascuna serie di 4 arrays. Brian fece similmente con i rimanenti 8 ratti ed 8 arrays (Figura 10.1)

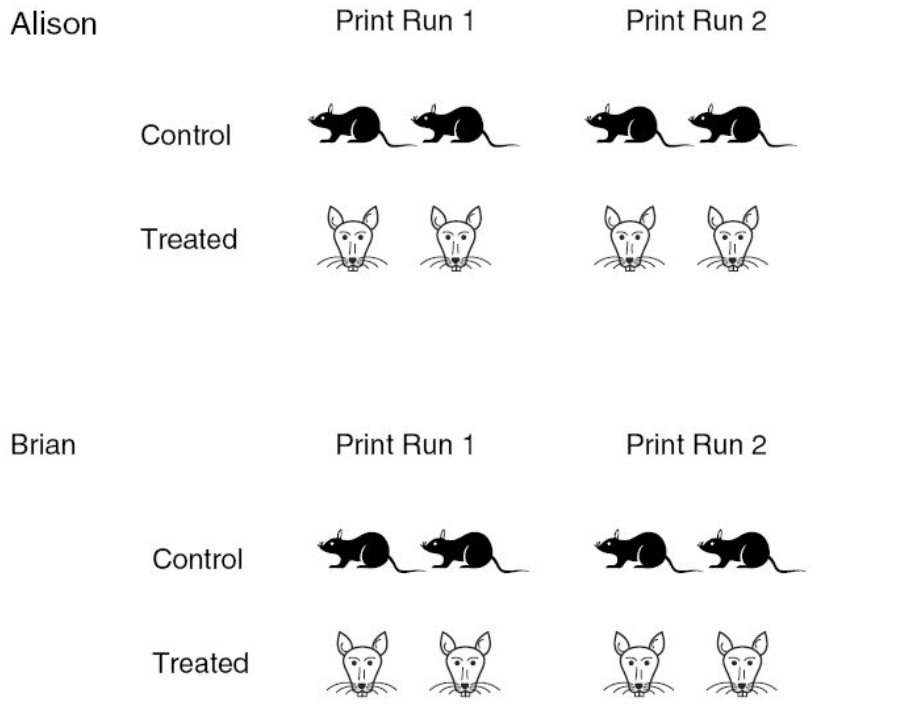


Figura 10.1: Progetto con Blocco Bilanciato. Vi sono 16 ratti che vengono usati in questo esperimento, 8 trattati con la tossina benzo(a)pirene, ed 8 trattati con una sostanza di controllo. I campioni verranno ibridizzati a due lotti di 8 arrays che derivano da due diversi cicli di produzione e sono coinvolti due ricercatori, Alison e Brian. Alison e Brian trattano ciascuno 4 ratti con la tossina e 4 ratti con la sostanza di controllo; essi preparano i campioni ed ibridizzano 2 campioni per ciascuno dei ratti trattati con tossina e con sostanza di controllo a 2 arrays provenienti da ciascuna delle fasi di produzione 1 e 2. Questo è un esempio di progetto bilanciato, che massimizza la potenza dell'esperimento.

Disegno Sperimentale 3

Otto ratti sono randomicamente allocati ad Alison; similmente, 4 arrays da ciascuno dei lotti sono allocati, ancora randomicamente, ad Alison. Quattro preparazioni di benzo(a)pirene e quattro preparazioni del compound di controllo vengono date ad Alison in modo che ella non conosca l'identità di alcuna preparazione. Gli arrays sono presistemati per Alison cosí che ella ibridizzerá due ratti trattati e due ratti di controllo a quattro arrays da ciascun batches, con allocazione casuale. Brian si comporta in modo simile.

Benché le imperfezioni dei primi due progetti possano apparire ovvie a molti ricercatori, è rimarchevole come parecchie storie che io ho sentito sugli esperimenti con i microarrays siano conformi al progetto 1. Il problema con il progetto 1 - che viene risolto con il progetto 2 - è la confusione; il problema con il progetto 2 - che è risolto nel progetto 3 - è la polarizzazione (bias).

Confondimento e Blocco

Supponiamo che venga usato il progetto 1 per l'esperimento, e che si faccia uso dell'analisi statistica (per esempio il *t*-Test o il test bootstrap descritti nel capitolo 7). Supponiamo che l'analisi statistica sia applicata per identificare i geni che sono up-regolati o down-regolati nel trattamento del gruppo relativo al gruppo di controllo. Noi vorremo essere in grado di poter dire quali di questi geni sono up - o down-regolati come effetto della epatotossina benzo(a)pirene. Di converso, la differenza osservata nel'espressione del gene potrebbe avere come causa il fatto che Alison e Brian abbiano gestito i campioni in modo differente, e quindi potrebbe non essere correlata alla tossina. Alternativamente, le differenze osservate nella espressione del gene potrebbero essere causate dalle differenze degli arrays nei due lotti e potrebbero, ancora, non essere correlate alla tossina. Con questo progetto sperimentale, noi non possiamo sapere quale dei tre fattori - trattamento, ricercatore o lotto - sia responsabile delle differenze nella espressione del gene. Diciamo che questi fattori sono confusi.

Il problema delle variabili confuse viene risolto per mezzo di una tecnica denominata **blocco.** Il progetto dell'esperimento 2 è un esempio di esperimento bloccato. In questo esempio vi sono due fattori bloccanti: lo sperimentatore ed il ciclo di produzione. Ciascuno degli 8 arrays è allocato in modo uguale tra i due fattori bloccanti (figura 10.1). Pertanto, se vi è una significativa differenza nel'espressione del gene tra i ratti trattati ed i ratti di controllo, è possibile attribuire la differenza al trattamento, e non può essere dovuta al ricercatore o al ciclo di produzione.

Questo è anche un esempio di progetto bilanciato. Un progetto sbilanciato può avere un numero diseguale di ratti nei gruppi di controllo o di trattamento, oppure un numero disuguale dei due gruppi allocati ad Alison ed a Brian. I progetti bilanciati sono molto più potenti dei progetti sbilanciati; noi discuteremo il significato di *"potenza"* nel paragrafo 10.3.

È importante notare che la fase *running* dell'esperimento in modalità bilanciata e bloccata non aggiunge nessun costo extra e nessun tempo addizionale all'esperimento, ma fa una differenza molto marcata in relazione al modo in cui si possono interpretare i risultati.

Polarizzazione, Randomizzazione ed Acciecamento

Il progetto sperimentale 2 soffre di un problema conosciuto come bias (polarizzazione). Quando Alison scelse i ratti, ella li poteva scegliere in modo che fossero in qualche modo simile: essi potevano apparire in pieno vigore, oppure molto docili, oppure in qualche altro modo. Non c'é nessun suggerimento sulla improprietà dell'operato di Alison: ogni persona fa la sua scelta inconscia senza accorgersene. Pertanto, avendo Alison scelto i ratti che ella usa, introduciamo una potenziale variabilitá tra i due gruppi (di ratti) usati dai due ricercatori. Il progetto sperimentale 3 usa la randomizzazione per rimuovere questa polarizzazione assegnando randomicamente i ratti ai due ricercatori.

Vi è una seconda sorgente di polarizzazione nel progetto sperimentale 2. Se Brian ed Alison sapessero a quali ratti essi stanno dando la tossina benzo(a)pirene, ed a quali stanno dando la sostanza di controllo, uno od entrambi (Alison e Brian) potrebbero comportarsi diversamente circa il modo in cui trattano i due gruppi di ratti. Ancora, non vi è alcun suggerimento di improprietà: vi possono essere fattori puramente inconsci che fanno in modo, per esempio, che il trattamento operato da Brian con la tossina, sia fatto con maggiore cura rispetto a quello dei ratti di controllo.

Il progetto sperimentale 3 usa **il blinding** (accecamento) per evitare questa polarizzazione: Alison e Brian non pongono attenzione su quali ratti sono trattati con la tossina e quali ratti sono trattati con la miscela (compound) di controllo.

Come risultato, essi possono trattare entrambi i gruppi di ratti allo stesso modo.

10.3 Scelta delle Tecnologie e Sistemazione dei campioni

I problemi delle variabili confuse e della polarizzazione sono sempre presenti in tutti progetti sperimentali, e non soltanto negli esperimenti con i microarrays.

Questo paragrafo discute tre problemi associati in special modo agli esperimenti con microarays:

- É meglio usare gli arrays Affimetrix, oppure sistemi di arrays a due colori?
- Se si usano sistemi di arrays a due colori, è meglio usare un campione di riferimento?
- Se si usano sistemi di arrays a due colori, quale è la migliore sistemazione dei campioni sui vetrini?

Non vi è una risposta universalmente valida a queste domande. Noi considereremo tre tipi di esperimenti e mostreremo come le considerazioni in ciascun caso portino a differenti conclusioni. Vi sono anche parecchi fattori che non possono essere determinati dalla statistica; per esempio, se la scelta degli array Affimetrix, oppure dei microarray a due colori, possa essere fatta in considerazione delle facilities (facilitazione in termini di attrezzature, disponibilità ecc.) disponibili in laboratorio. In modo simile, discuteremo se l'uso dei due diversi tipi di array in un progetto che richiede 20 o 40 array, possa essere dettato da condizioni economiche. Di converso, vi sono ragioni statistiche dalle quali scaturisce che alcuni di questi progetti siano migliori di altri. Questi ed altri aspetti verranno focalizzati alla fine del paragrafo.

Esempio 10.2: Carcinoma epatocellulare

Sono stati prelevati campioni da tessuti malati e da tessuti sani in pazienti affetti da carcinoma epatocellulare ed ibridizzati ai microarray. Noi vorremmo essere in grado di

identificare geni che sono up-regolati o down-regolati nel carcinoma epatocellulare relativamente al tessuto sano.

Progetto Sperimentale 1

Sono stati usati quaranta microarray. Qualsiasi campione di tessuto sano e malato è preparato con Cy5 (rosso). Un campione di riferimento di linee cellulari rilevanti è marcato con Cy3 (verde). Ciascun array è ibridizzato con un campione di fegato marcato con Cy5 e con il campione di riferimento marcato con Cy3 (Figura 10.2a).

Progetto Sperimentale 2

Sono usati 40 arrays Affimetrix. Ciascun array è ibridizzato con un campione differente (Figura 10.2b).

Progetto Sperimentale 3

Sono usati 20 microarrays; su ciascun array, il campione di tessuto sano è ibridizzato con Cy3, ed il campione con il tumore, proveniente dallo stesso paziente, è ibridizzato con Cy5 (Figura 10.2c).

Progetto Sperimentale 4

Sono usati 20 microarrays. Su 10 arrays è ibridizzato il campione con tessuto sano con Cy3 ed il campione di tessuto con tumore è ibridizzato con Cy5. Sugli altri dieci array, il campione con tessuto sano è ibridizzato con Cy5 ed il campione con tessuto tumorale è ibridizzato con Cy3 (Figura 10.2d).

Progetto Sperimentale 5

Sono usati 40 microarrays. Ciascun campione - sia di tessuto sano che di tessuto malato - è marcato due volte, una volta con Cy3 ed una volta con Cy5. I campioni di tessuto sano e malato di ciascun paziente sono ibridizzati ai due arrays, una volta con il tessuto sano in Cy3 e con il tessuto tumorale in Cy5, ed una volta con tessuto sano in Cy5 e con tessuto malato in Cy3 (Figura 10.2e).

Il primo punto da tenere in considerazione è che noi abbiamo descritto cinque differenti progetti sperimentali per un semplice esperimento con microarrays. Vi sono parecchie scelte su come condurre questo esperimento, e le scelte che si faranno avranno un impatto sull'interpretazione dei dati. Vi sono molte differenze ovvie tra i diversi progetti: i progetti 1, 3, 4, e 5 usano arrays a due colori, mentre il progetto 2 usa gli array Affimetrix; i progetti 1,2 e 5 usano 40 arrays, mentre i progetti 3 e 4 usano 20 arrays. Il progetto 1 include un campione di riferimento. Quale è il miglior modo di condurre questo esperimento? Questo esempio è un esperimento accoppiato e quindi richiede una analisi a coppie dei dati (Paragrafo 7.1). Ricordiamo: si stanno comparando due campioni dello stesso paziente con lo scopo di identificare i geni che sono up-regolati e down-regolati, ed i due campioni provenienti dai due pazienti hanno

una ovvia relazione l'uno con l'altro. In base a ciò, appare chiara la ragione di usare un array a due colori e di ibridizzare i due campioni dello stesso paziente agli stessi arrays.

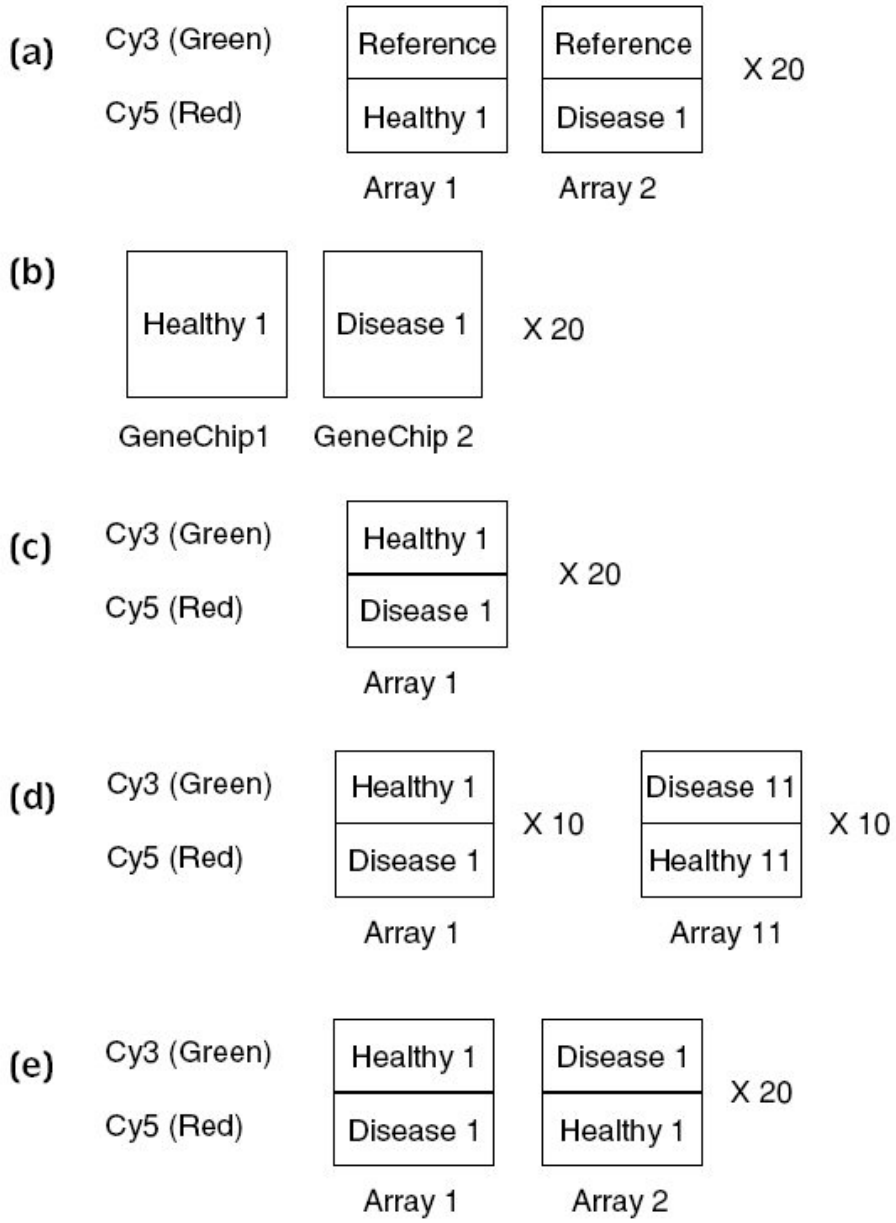


Figura 10.2: Progetti sperimentali per un semplice esperimento con microarray. In un esperimento condotto per identificare geni espressi differenzialmente in 20 pazienti affetti da epatocarcinoma cellulare, sono stati prelevati tessuti sani e tessuti malati, e quindi sono stati preparati campioni ibridizzati ai microarray. Perfino questo semplice esperimento implica cinque possibili progetti sperimentali. **a)** Ciascuno dei 40 campioni è ibridizzato con microarray separati, con un campione di riferimento separato, ibridizzato al secondo canale. **(b)** Ciascuno dei 40 campioni è ibridizzato ad un diverso GeneChip Affimetrix. **(c)** I due campioni di ciascun paziente sono ibridizzati allo stesso array, con il campione del tessuto sano in Cy3 ed il campione del tessuto malato in Cy5; sono necessari –in questo caso-- soltanto 20 arrays. **(d)** come in (c), ad eccezione di 10 pazienti che hanno il tessuto sano in Cy3 ed il tessuto malato in cy5, mentre gli altri 10 pazienti hanno il tessuto sano in Cy5 ed il tessuto malato in Cy3. **(e)** I campioni di ciascun paziente sono marcati due volte: una volta con Cy3 ed una volta con Cy5. Questi campioni sono quindi ibridizzati a due array in un esperimento di interscambio del fluorocromo.

Stima della variabilità

Si può stimare la variabilità dei differenti progetti sperimentali con un semplice calcolo basato sul modello log-normale introdotto nel capitolo 6. Se il coefficiente di variabilità del segnale di ibridizzazione sull'array è *v*, allora la varianza del logaritmo del segnale σ^2 è data dalla seguente equazione:

$$\sigma^2 = \ln(v^2 + 1)$$

Equazione 10.1

L'Equazione 10.1 è identica all'equazione 6.1. Se, per esempio, il coefficiente di variabilitá é del 30%, allora la varianza è 0.086, e la deviazione standard del logaritmo del segnale è 0.29. Si noti che questo è il logaritmo naturale; per calcolare la deviazione standard in base 2, si deve dividere la deviazione standard con il ln(2). In questo caso, la deviazione standard del log in base 2 sarebbe 0.42.

Con il progetto sperimentale 1 (campione di riferimento), il log del rapporto della espressione del gene tra i due campioni è calcolata indirettamente calcolando i logaritmi dei rapporti dei campioni del tessuto sano e del tessuto malato, rispetto ai campioni di riferimento, e quindi sottraendo questi due rapporti logaritmici. Poiché vi sono quattro campioni implicati nel calcolo, vi sono quattro contributi della varianza (sigma quadro) alla varianza totale, che è 4 x 0.086 = 0.344. Così la deviazione standard del logaritmo del rapporto è la radice quadrata di 0.344 = 0.59.

Con il progetto sperimentale 2 (Affimetrix) e 3 (stessa ibridizzazione dell'array), il rapporto logaritmico è calcolato direttamente. Nel progetto 2 calcoliamo il rapporto logaritmico dei campioni come logaritmo del rapporto dei due segnali Affimetrix, e nel progetto 3, calcoliamo il logaritmo del rapporto dei segnali rosso e verde dell'array. Vi sono soltanto due contributi della varianza (sigma quadro) alla varianza totale. Così che la varianza totale è 0.172 e la deviazione standard del rapporto logaritmico è 0.41.

Pertanto, i progetti sperimentali 2 e 3 hanno errori più piccoli e quindi entrambi sono migliori rispetto al progetto sperimentale 1 in questo tipo di esperimento.

Questo calcolo non è esatto; abbiamo assunto che la variabilitá delle ibridizzazioni tra gli arrays sia uguale alla variabilitá delle ibridizzazioni dello stesso array. Nel capitolo 6 abbiamo visto che questo non è vero. La variabilità tra gli arrays è generalmente più grande rispetto alla variabilitá dei segnali sullo stesso array; ciò fa in modo che questo esperimento sia meglio sviluppato su array a due colori dove i due campioni provenienti da ciascun paziente possono essere ibridizzati sullo stesso array che sugli array Affimetrix, dove i due campioni di ciascun paziente devono essere allocati su arrays differenti.

Confusione e scambio di colori

Il progetto sperimentale 3 non è, per ulteriori ragioni, il miglior progetto per questo esperimento. C'è un problema: tutti i campioni del tessuto sano sono marcati con Cy3, e tutti i campioni del tessuto malato sono marcati con Cy5. Se, nel corso dell'analisi, noi vediamo un gene che è differenzialmente espresso, ciò potrebbe significare uno stato di malattia, oppure potrebbe significare una incorporazione differenziale dei fluorocromi. Con il progetto sperimentale 3, la marcatura è confusa con il fattore che ci interessa (tessuto malattia/salute), e non si può dire quali di questi fattori è responsabile dell' espressione differenziale del gene che andiamo osservando.

Nei progetti sperimentali 4 e 5 questo problema è stato risolto. Il progetto sperimentale 4 è un progetto bilanciato bloccato. I fluorocromi rosso e verde sono variabili bloccate, e quindi è possibile determinare i geni che sono differenzialmente espressi nel tessuto malato. Il progetto sperimentale 5 è quello conosciuto come progetto *full-factorial* (totalmente-fattoriale), poiché ciascun paziente ha avuto ciascun campione ibridizzato due volte: una volta per ciascuno fluorocromo. Ci sono due vantaggi con il progetto sperimentale 5 rispetto al progetto sperimentale 4, ma anche uno svantaggio. Il primo vantaggio del progetto 5 consiste nel fatto che ci siano due misure del rapporto logaritmico per ciascun paziente, usando lo stesso numero di arrays del progetto sperimentale 1 (il campione di riferimento). Questa è una sorta di replicazione tecnica che riduce la deviazione standard dei rapporti logaritmici misurati, di un fattore radice di due; per esempio, con un coefficiente di variabilitá del 30%, la deviazione standard del rapporto logaritmico nel progetto sperimentale 5 dovrebbe essere di 0.29, in confronto a 0.41 del progetto sperimentale 4, e 0.59 del progetto sperimentale 1.

Il secondo vantaggio consiste nel fatto che i dati possono essere analizzati immediatamente con il t-test o con il bootstrap t-test, mentre il progetto sperimentale 4 richiede una analisi ANOVA molto più complicata, ed un bootstrap molto più complesso per ottenere *p*-values (valori del *p*-values) (Paragrafo 7.6)

Lo svantaggio ovvio del progetto sperimentale 5 è che esso richiede il doppio di arrays rispetto al progetto sperimentale 4 ed il doppio di reazioni di marcatura rispetto a qualsiasi altro progetto. Il progetto che verrà scelto dipenderà, quindi, dalle risorse finanziarie di cui il laboratorio dispone.

Esempio 10.3: Linfoma con cellule B

I campioni sono prelevati da pazienti affetti da linfoma a cellule B e sono ibridizzati ai microarray. Lo scopo dell'esperimento è l'identificazione di sottogruppi di pazienti clinicamente rilevanti usando l'analisi del cluster, e quindi costruire un modello di classificazione per differenziare i vari sottogruppi.

Progetto Sperimentale 1

Sono preparati campioni di 30 pazienti e marcati con Cy3, ed i campioni dei restanti 30 pazienti marcati con Cy5. Questi sono ibridizzati a 30 diversi array a due colori (Figura 10.3a).

Progetto Sperimentale 2

I campioni di ciascun paziente sono preparati e marcati con Cy3 ed ibridizzati con 60 differenti array a due colori; un campione di riferimento universale è ibridizzato in Cy5 con ciascun array (Figura 10.3b).

Progetto Sperimentale 3

I campioni di ciascun paziente sono preparati ed ibridizzati con 60 differenti array Affimetrix (figura 10.3c).

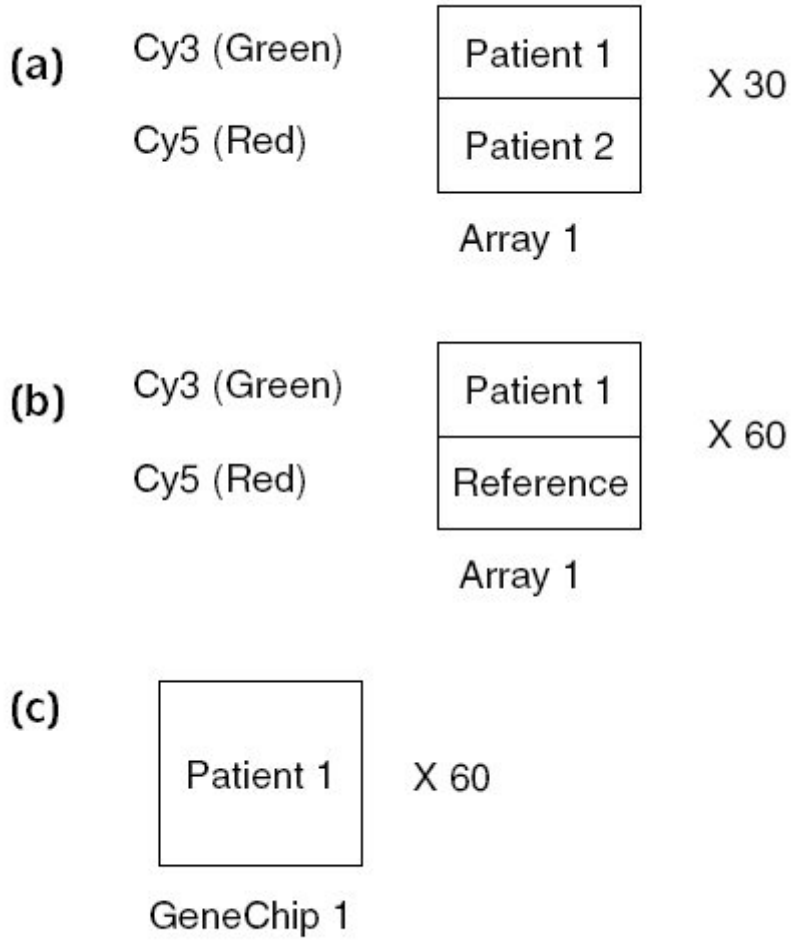


Figura 10.3: progetto sperimentale per lo studio del linfoma. Sono stati prelevati campioni da 60 pazienti affetti da linfoma diffuso a cellule grandi. Lo scopo è quello di identificare sottogruppi clinicamente rilevanti di pazienti usando l'analisi del cluster, e quindi costruire un predittore per una analisi differenziata attraverso le classi. Vi sono tre possibili progetti sperimentali: **(a)** i campioni di 30 pazienti sono marcati con Cy3, ed i campioni degli altri 30 pazienti sono marcati con Cy5. Questi sono quindi ibridizzati a differenti arrays a due colori. **(b)** i campioni di tutti i 60 pazienti sono tutti marcati con Cy3. Questi sono ibridizzati con Cy3 ed ibridizzati con 60 arrays tutti separati. Un campione di riferimento universale è ibridizzato con Cy5 con ciascuno array. **(c)** I campioni di ciascun paziente sono ibridizzati con 60 array Gene-Chip Affimetrix.

Il progetto sperimentale 1 non è un buon progetto. Se vogliamo applicare i metodi di clustering del capitolo 8, dobbiamo essere in grado di confrontare ciascun campione l'un l'altro su una base comune. D'altra parte, questo progetto non ci permette di fare ciò. Benché le coppie di campioni scorrelati che sono stati ibridizzati sullo stesso array possano essere comparati facilmente, è molto difficile fare un confronto tra due campioni ibridizzati su array differenti, specialmente se questi sono marcati con differenti fluorocromi. In tal modo, benché si potrebbe essere indotti nella tentazione di usare un progetto che richieda metà del numero degli array usati nei progetti 2 e 3, i dati derivanti da questo esperimento non si presterebbero ad essere analizzati naturalmente. Il progetto sperimentale 2, d'altra parte, è molto più adatto alle analisi che si vogliono fare. Ogni campione può essere normalizzato ad un campione di

riferimento e quindi paragonato in modo significativo con gli altri in una analisi di clustering o di classificazione.

Il progetto sperimentale 3 è anche un buon progetto. L'uniformità della piattaforma Affimetrix rende significativo il confronto tra i campioni; ogni array "buio" può essere normalizzato usando la normalizzazione degli arrays descritta nel paragrafo 5.4.

Esempio 10.4: Serie Temporali

Il lievito gemmante può riprodursi per via sessuale attraverso la produzione di cellule aploidi mediante un processo denominato sporulazione. Il lievito fu messo in un mezzo sporulante ed i sette campioni furono presi in successione temporale dall'inizio della sporulazione. Siamo interessati ad identificare i geni che mostrano profili simili nel corso del tempo.

Progetto Sperimentale 1

I campioni provenienti dai sette punti temporali sono ibridizzati con sette array Affimetrix (figura 10.4a).

Progetto Sperimentale 2

I campioni provenienti dai sei punti temporali a partire dall'istante zero sono preparati e marcati con Cy3. Un campione piú grande a partire dall'istante zero è preparato e marcato con Cy5 come campione di riferimento[1]. I campioni sono ibridizzati a 6 arrays con ciascun punto temporale nel canale Cy3, e con il campione al tempo zero nel canale Cy5 (figura 10.4b).

Progetto Sperimentale 3

I campioni dei sette punti temporali sono marcati due volte ciascuno: una volta con Cy3 ed una con Cy5. Gli arrays sono ibridizzati con sette arrays come mostra la figura 10.4c. Questo progetto é conosciuto come *progetto ad anello chiuso* (loop design).

[1] I precedenti esperimenti evolventi nel tempo usavano il campione al tempo zero come campione di riferimento. In tempi piú recenti, i ricercatori stanno impiegando una pratica migliore usando una miscela di campioni da tutti i punti temporali come un campione di riferimento comune. Questo ha il vantaggio di assicurare che vi siano segnali nel campione di riferimento provenienti da tutti i geni che sono espressi allo stesso punto nel corso del tempo

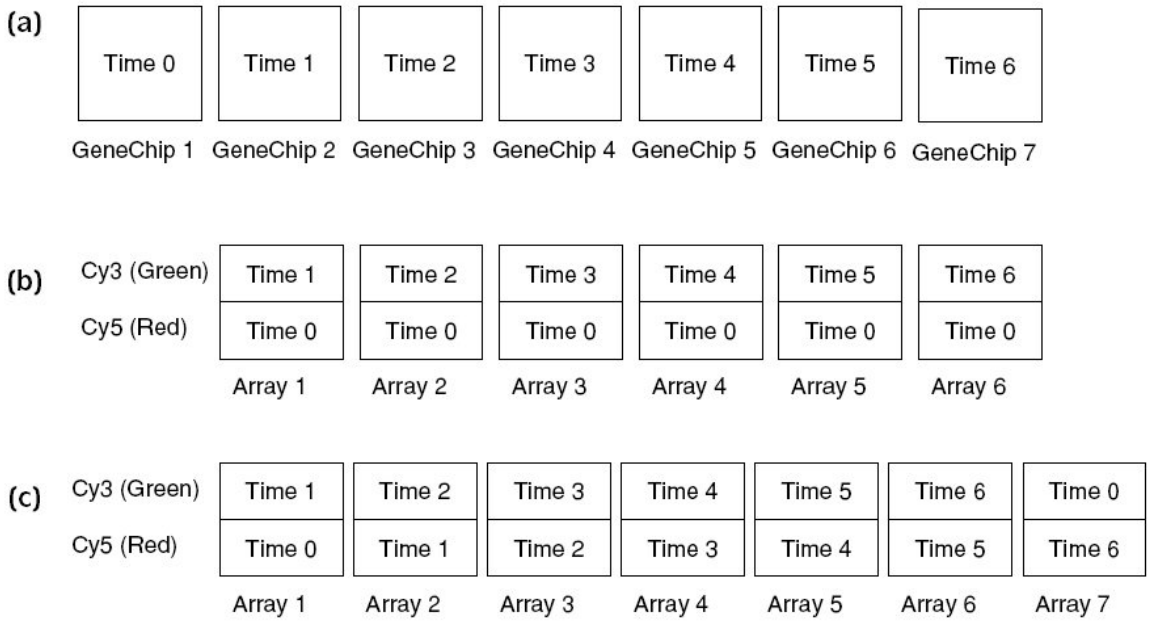


Figura 10.4: Progetto sperimentale per lo studio di una serie temporale. L'inoculo di lievito è trattato con un mezzo sporulante: i campioni sono presi a sette differenti punti temporali ed ibridizzati con i microarray. Vi sono tre progetti sperimentali piuttosto comuni. **a)** I campioni provenienti da ciascun punto temporale sono ibridizzati con sette differenti GeneChip Affimetrix. **b)** Il campione relativo al tempo zero è utilizzato come campione di riferimento, marcato con Cy5 e ibridizzato con tutti gli arrays. I campioni provenienti dagli altri sei punti temporali sono marcati con Cy3 ed ibridizzati con sei differenti arrays a due colori. **c)** I campioni provenienti da tutti e sette i punti temporali sono marcati due volte: una volta con Cy3 ed una volta con Cy5. Questi sono ibridizzati con gli arrays secondo il pattern mostrato. Questo procedimento è conosciuto come loop design (progetto ad anello).

Il progetto sperimentale 1 presenta un serio problema, che dovrebbe essere presente anche quando si sviluppi questo tipo di esperimento usando un microarray ad un solo colore, oppure quando si usino campioni marcati radioattivamente su filtri di nylon. La stimolazione di una coltura di cellule può portare, verosimilmente, a cambi globali della espressione del gene per tutto il corso di evoluzione delle serie temporali. Supponiamo di considerare un particolare array, corrispondente ad un particolare punto temporale, che presenti una luminosità più elevata rispetto agli altri. Ciò potrebbe avere due interpretazioni. Primo, può darsi che si tratti di un artefatto sperimentale derivante dalla ibridizzazione differenziale; secondo, potrebbe trattarsi di una generale espressione del gene piú elevata a quel particolare punto temporale. (Figura 10.5).

Se noi usiamo la tecnologia Affimetrix, od un differente sistema a singolo colore, questi due fattori sono confusi e nessuna analisi puó condurre a ricostruire la vera situazione. Benché ci possa essere la tentazione di applicare la normalizzazione *tra gli array* (paragrafo 5.4), se ciò venisse fatto, sarebbe scorretto poiché rimuoverebbe dall'analisi tutta la informazione circa i cambiamenti globali nell'espressione del gene (figura 10.5b). I progetti sperimentali 2 e 3 risolvono questo problema. Con il progetto sperimentale 2, ciascun campione è normalizzato rispetto al campione al tempo zero, cosí che le misure sono rapporti logaritmici di punti temporali relativi al tempo zero.

L'assunzione è che se un array è particolarmente luminoso, esso sarà luminoso per entrambi i campioni (figura 105.a), e così i rapporti logaritmici saranno liberi da artefatti (Figura 10.5b). Il progetto sperimentale 3 ha il vantaggio, rispetto al progetto sperimentale 2, che vi sono due misure indipendenti per ciascun campione usando lo stesso numero di array. Di converso, vi sono due svantaggi. Il primo è che esso

richiede una analisi ANOVA piú complessa per essere in grado di confrontare tutti i campioni su tutti gli array (Paragrafo 7.6), in contrasto con il progetto sperimentale 2, che può essere rapidamente analizzato senza far uso di ANOVA. Il secondo è che se un singolo array dovesse venire male, esso pregiudicherebbe l'intera analisi. Con il progetto sperimentale 2, i dati provenienti da un array fuori uso possono essere omessi, e gli altri dati dei punti possono continuare ad essere usati.

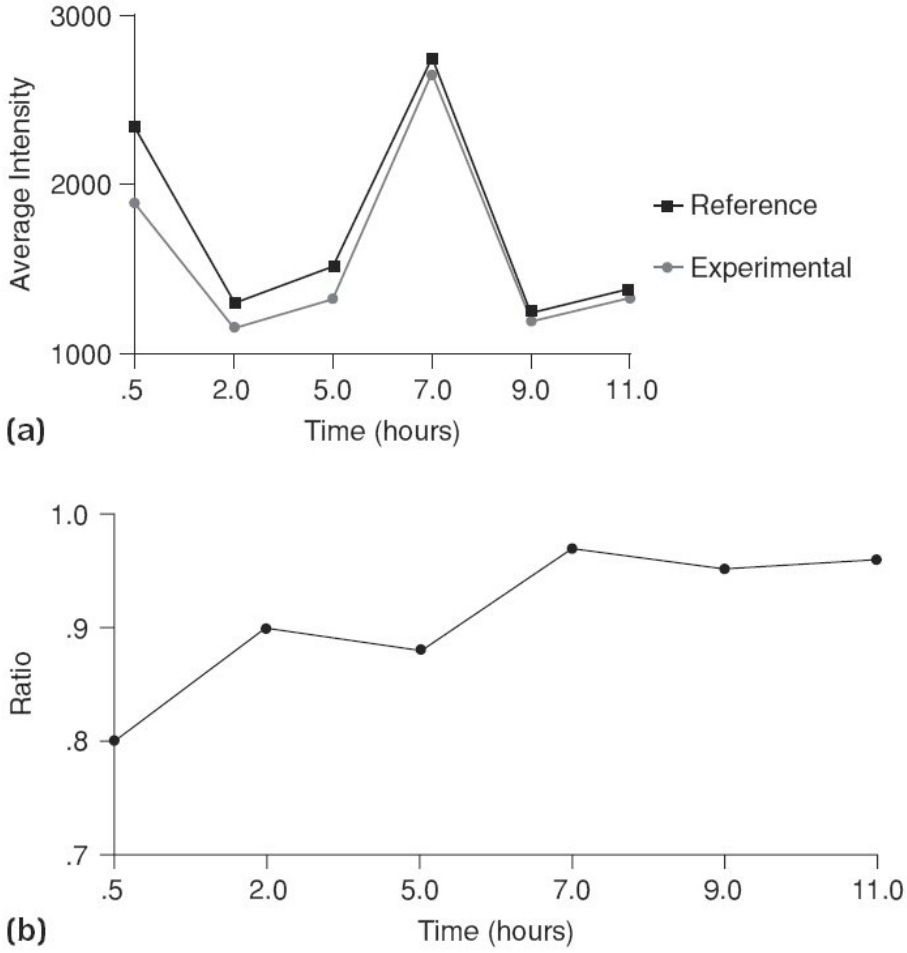


Figura 10.5: Il problema della normalizzazione in esperimenti a singolo-colore in funzione del tempo. Un innesto di lievito viene posto in un mezzo sporulante ed i campioni sono presi al tempo zero, e dopo 30 minuti, 2, 5, 7 , 9 ed 11 ore. L'esperimento è progettato in accordo al progetto sperimentale 2, su arrays a due colori, con campione preso al tempo zero usato come campione di riferimento. **(a)** intensità media per tutti i geni per entrambi i campioni: sperimentale e di riferimento. L'array dopo 7 ore è molto più luminoso che in qualsiasi altro punto. Se questo esperimento fosse stato sviluppato senza il campione di riferimento, per esempio su un piattaforma Affimetrix, non sarebbe stato possibile determinare se questo segnale così alto avesse rappresentato la massima induzione genica all'istante temporale di 7 ore, oppure si fosse trattato semplicemente di un array luminoso. Infatti, il segnale sul campione di riferimento è anch'esso elevato, suggerendo che questo sia un array luminoso. **(b)** media del rapporto logaritmico di tutti i geni sull'array rispetto al campione di riferimento. Il rapporto logaritmico incrementa con il tempo fino a 7 ore, fatto che indica che l'espressione dell'intero genoma incrementa in questo intervallo di tempo. Pertanto, la normalizzazione per mezzo del segnale medio dell'array non sarebbe corretta poiché questa informazione verrebbe perduta. In conclusione, è impossibile analizzare questi dati senza il supporto del campione di riferimento oppure senza l'ausilio di un progetto loop.

10.4 Quanti Replicati

Questa è la terza considerazione sul progetto sperimentale e una delle domande più frequenti circa gli esperimenti con i microarrays: quanti replicati? La risposta a questa domanda dipende da un certo numero di fattori; piuttosto critici sono sia il tipo di esperimento che si sta pianificando, sia l'analisi con la quali i dati verranno analizzati.

In questo paragrafo, considereremo i metodi per stimare il numero di replicati necessari per esperimenti che rivelino l'espressione differenziale dei geni (Capitolo7).

Per molti scopi pratici, il calcolo del numero di replicati di questo tipo di esperimenti è adeguato anche per quegli esperimenti in cui verrà applicata la "Cluster Analysis" (Capitolo 8) o l'analisi delle classificazioni (Capitolo 9). Questo perché il numero di replicati può essere usato per determinare la soglia di espressione differenziale che vogliamo fissare per includere un gene in un'analisi di clustering o di classificazione.

Il modo classico per stimare il numero di replicati necessari ad un esperimento è **l'analisi della potenza**. Il concetto di potenza è correlato molto da vicino ad un altro concetto: la confidenza, che abbiamo accennato nel capitolo 7.

Confidenza e Potenza

La **Confidenza** di un test statistico è la probabilità di non avere un risultato falso positivo. Per dirla in un'altra maniera, è la probabilità con la quale si afferma che il gene non si sia espresso in modo differenziale, quando il gene non si é realmente espresso in modo differenziale. Da un punto di vista statistico, essa (la confidenza) rappresenta la probabilità di accettare l'ipotesi nulla quando l'ipotesi nulla è vera. Gli statistici qualche volta si riferiscono a risultati falsi positivi come errori di Tipo I. Gli errori di Tipo I sono generalmente controllati esplicitamente quando si seleziona un livello di significatività per il test statistico. Per esempio, quando si seleziona un test statistico con una soglia di significatività dell'1%, si sta selezionando un confidenza del 99%.

Negli esperimenti con i microarrays, la confidenza deve essere aggiustata per tenere conto della molteplicità dei test (paragrafo 7.5).

La **Potenza** di un test statistico è la probabilità di non avere risultati falsi negativi.

Questa è la probabilità con cui si afferma che un gene sia differenzialmente espresso quando un gene è realmente differenzialmente espresso. Da un punto di vista statistico, essa rappresenta la probabilità di reiezione dell'ipotesi nulla, quando l'ipotesi nulla è falsa. Gli statistici qualche volta si riferiscono ai risultati falsi negativi come errori di Tipo II. Gli errori di Tipo II non possono essere controllati esplicitamente, ma sono controllati implicitamente per mezzo del progetto sperimentale. In particolare, la potenza di un esperimento dipende in modo critico dal numero dei replicati usati, perciò la scelta del numero di replicati è determinato dalla potenza che si vuole raggiungere nell'analisi.

Gli errori di Tipo I e di Tipo II sono riassunti nella tabella 10.1. Quando si sceglie una soglia di significatività, tale soglia impone una scelta circa il bilanciamento tra potenza e confidenza dell'analisi: una soglia di significatività più stringente fornisce un piú alto valore di confidenza ma riduce la potenza, e, inversamente, una soglia meno stringente di significatività fornisce meno confidenza ma aumenta la potenza[2].

[2] Una domanda comune è quale sia peggiore: un errore di Tipo I o un errore di Tipo II. Non vi é una risposta statistica a questa questione: in situazioni differenti , gli errori Tipo I o Tipo II possono essere meno preferiti. La Statistica ci fornisce un tool con il quale si possono misurare i ratei di questi errori, ma il come bilanciare i due tipi di errore è soggettivo. Per esempio, se lo scopo dell'esperimento con microarray è quello di identificare nuovi target o la scoperta di medicinali, e se si intende spendere molto denaro per caratterizzare questi target, allora può essere più importante non avere errori di Tipo I, poiché un falso positivo vuol dire un costoso fallimento. D'altra parte, se il

TABLE 10.1: Type I and Type II Errors

	True Situation	
Our Decision	**Not Differentially Expressed**	**Differentially Expressed**
Not significant	Correct	Type II error
Significant	Type I error	Correct

***Nota*:** Vi sono quattro possibili risultati in un test statistico. I due risultati corretti si verificano sia se il gene non è differenzialmente espresso, e noi diciamo che non è significativo, oppure se il gene è differenzialmente espresso, e noi diciamo che è significativo. Vi sono, altresì, due possibili risultati sbagliati: un errore di tipo 1 ha luogo quando il gene non è differenzialmente espresso e l'analisi conclude, invece, che è significativo; un errore di Tipo II ha luogo, invece, quando il gene è differenzialmente espresso e l'analisi conclude che non è significativo. La confidenza dell'analisi è la probabilità di non avere un errore di Tipo I; la potenza dell'analisi è la probabilità di non avere un errore di Tipo II. La potenza è controllata scegliendo un numero appropriato di replicati biologici.

La ragione per cui viene usata la potenza dell'analisi per determinare il numero di replicati in un esperimento, è perché la potenza di un test statistico dipende dai seguenti fattori:

- Il numero di replicati
- Il tipo di analisi (a coppie o non a coppie)
- La differenza in media che stiamo cercando di rivelare (che è il rapporto logaritmico)
- La deviazione standard della variabilità della popolazione
- La soglia di significatività del test

Pertanto, noi stimiamo il numero di replicati necessari dalla conoscenza degli altri parametri e con una potenza desiderata predeterminata. Prima di mostrare come funziona questa analisi, discuteremo ciascun parametro in dettaglio.

Tipi di Replicati

Gli esperimenti con microarray possono essere replicati a molti livelli differenti. Fondamentalmente, vi sono due tipi di replicati: replicati **biologic**i e replicati **tecnici**. I replicati biologici sono replicati presi a livello della popolazione che si sta studiando.

Nell'esempio 10.2, dove i campioni sono presi da 20 pazienti affetti da carcinoma epatocellulare, ciascun paziente è un replicato biologico. Nel'analisi di potenza, i replicati verranno intesi sempre come replicati biologici. Ciò avviene perché l'analisi che noi intendiamo sviluppare, fornisce una inferenza statistica sulla popolazione dalla quale i replicati provengono (paragrafo 7.1). Per questo è necessario includere sufficienti replicati biologici per essere certi che gli effetti che vediamo possano essere generalizzati alla popolazione (e.g., la popolazione di pazienti affetti da carcinoma epatocellulare). I replicati tecnici, d'altro canto, sono presi a livello di apparato sperimentale. Lo scopo dei replicati tecnici è di tener conto della variabilità nel setup dell'esperimento (impostazione dei parametri, volumi, tempi di campionamento e quant'altro dell'apparato sperimentale); se l'esperimento fosse di qualità sufficiente,

microarray viene usato come tool diagnostico per un programma di salute pubblica per un cancro fatale, come per esempio il carcinoma alla mammella, allora puó essere importante non avere errori di Tipo II. Un falso negativo potrebbe risultare in un paziente che sviluppa un tumore fatale che poteva essere curabile sea rivelato agli esordi.

allora non ci sarebbe affatto bisogno di replicati. I replicati tecnici possono essere a parecchi livelli:

- Spot replicati sull'array, che possono dare informazione su differenze nel printing o nell'ibridizzazione
- Replicati di array ibridizzati con lo stesso campione
- replicati di passaggi preparativi; per esempio, due marcature a fluorocromi invertiti

I replicati tecnici non possono essere considerati come campioni differenti sia nei calcoli della potenza che nell'analisi dei dati che sono prodotti. Invece, è pratica comune prendere la media aritmetica dei replicati tecnici per fornire una singola misura per l'individuo. Questo migliora nettamente l'affidabilità dei dati del microarray, poiché la variabilitá sperimentale, in una media di parecchi replicati tecnici, diminuisce.

Qualche utilizzatore di microarray adopera un pool di campioni provenienti da parecchi individui prima di ibridizzarli all'array. Per scopi di calcolo sia della potenza che di analisi dei dati, i campioni selezionati dal pool contano come un singolo individuo, poiché l'informazione circa la variabilità tra gli individui è andata perduta nel processo di pooling. Pertanto, negli esperimenti dove la variazione tra gli individui è importante (e.g., esperimenti sulle malattie dell'uomo), il pooling dovrebbe essere evitato per quanto possibile.

Tipo di Analisi

Nel capitolo 7 noi abbiamo considerato dati che erano sia a coppie che singoli. La potenza è differente nei dati a copia rispetto ai dati singoli: in generale, i test a coppie sono molto più potenti poiché le differenze individuali sono cancellate dal meccanismo di appaiamento. In questo paragrafo, discuteremo l'analisi per questo tipo di dati. È anche possibile sviluppare l'analisi della potenza per tipi di dati con struttura più complessa, come per esempio quelli richiesti da ANOVA; questa è statistica molto avanzata.

Differenza in Media

La potenza di un test statistico dipende dalla differenza in media che stiamo cercando di estrarre. Ne caso dei dati accoppiati, questa è la differenza tra la media dei dati e lo zero; nel caso di dati disaccoppiati, questa è la differenza tra i due gruppi. Nell'analisi del microarray, dove stiamo lavorando con dati registrati, la differenza in media viene traslata al rapporto logaritmico medio. Dovrebbe essere abbastanza intuitivo che è più difficile rivelare piccole differenze in media che differenze ampie.

Per esempio, è molto più difficile rivelare espressioni differenziali di 1.5 volte dei geni che espressioni differenziali di 3 volte. Di converso, è importante appezzare che noi non usiamo il rapporto di espressione differenziale come soglia per la rivelazione dei geni; stiamo semplicemente affermando che la potenza dei test di ipotesi descritti nel capitolo 7 dipende dal livello dell'espressione differenziale, oltre che da altri parametri.

Deviazione Standard

La potenza di un test statistico dipende anche dal livello di variabilità nella popolazione. In questi calcoli, noi facciamo l'assunzione che gli errori nelle misure di espressione del gene siano statisticamente distribuiti secondo una curva log-normale

(vedere capitolo 6). Questa assunzione è soltanto approssimativamente vera per la maggior parte di esperimenti con microarray; per questa ragione non raccomandiamo l'uso del *t*-Test per identificare i geni differenzialmente espressi (paragrafo 7.3). Di converso, le analisi della potenza sono soltanto una guida approssimata per la stima del numero di replicati e non sono una misura precisa; pertanto, le deviazioni dall'assunzione log-normale non sono un problema serio. Se si preferisce sviluppare delle analisi di potenza senza assumere che i dati siano distribuiti in modo log-normale, è possibile sviluppare l'analisi di potenza bootstrap; il lettore interessato è rimandato al libro sul bootstrapping elencato alla fine del capitolo 7.

Calcoli dell'analisi della potenza e Tabelle

Le formule per l'analisi della potenza sono alquanto complicate. Di converso, molti pacchetti statistici implementano l'analisi della potenza; in questo paragrafo noi mostreremo come si debba usare la funzione di analisi della potenza con il package R. Ciò presenta dei vantaggi che permettono allo sperimentatore di selezionare livelli di confidenza molto stringenti. Questa flessibilità è necessaria poiché negli esperimenti con i microarray usiamo livelli di confidenza molto alti (o soglie di significatività molto basse) con lo scopo di controllare il rateo di falsi positivi.

La funzione R e:

power.t.test(*n, delta, sd, sig.level, power, type, alternative*)

Equazione 10.2

Dove:

- *n* è il numero di replicati (in un e test a campione singolo) oppure grandezza del gruppo (in un test a due campioni);
- *delta* è la differenza in media che stiamo cercando di rivelare (che è il rapporto logaritmico);
- *sd* è la deviazione standard della variabilità della popolazione (calcolata usando l'equazione 10.2);
- *sig.level* è l soglia di significatività;
- *power* è la potenza desiderata;
- *type* può essere a singolo o a due campioni; a campione singolo è usato per analisi accoppiate, e a due campioni per analisi non accoppiate;
- *alternative* può essere ad una coda o a due code. (negli esperimenti con i microarray si presentano quasi sempre entrambi i casi per i geni up-regolati o down-regolati, così che - in generale - noi utilizziamo due code).

Per usare la formula, una delle variabili *n, delta, sd, sig.level* o *power* è omessa, e quindi la funzione calcola il valore della variabile omessa. Di solito o si omette *n* e si fornisce la potenza desiderata in modo che la formula restituisca il numero di replicati di cui necessitiamo, oppure si omette *power* e si fornisce il numero di replicati che si stanno studiando, in modo che la formula restituisca la potenza del nostro esperimento.

Nella Tabella 10.2, abbiamo usato il pacchetto R per calcolare le potenze con analisi a singolo o a due campioni per rivelare i geni espressi differenzialmente, con una significatività di 0.0001 (che darebbe un falso positivo su 10.000 geni del microarray), per una varietà di livelli di variabilitá della popolazione ed ampiezze dei gruppi. I due esempi che seguono mostrano come sviluppare una analisi di potenza per esperimenti specifici.

TABLE 10.2A: Power Analysis for Paired Test for 2-Fold Difference with $\alpha = 0.0001$

Num Reps	Population Coefficient of Variation									
	20%	25%	30%	35%	40%	45%	50%	60%	70%	80%
3	0.4%	0.2%	0.2%	0.1%	0.1%	0.1%	0.1%	0.1%	0.0%	0.0%
4	2.2%	1.2%	0.7%	0.5%	0.4%	0.3%	0.2%	0.1%	0.1%	0.1%
5	9.7%	4.8%	2.7%	1.6%	1.1%	0.8%	0.6%	0.3%	0.2%	0.2%
6	29.7%	14.8%	8.0%	4.7%	2.9%	2.0%	1.4%	0.8%	0.5%	0.4%
7	59.4%	33.7%	18.9%	11.0%	6.8%	4.4%	3.0%	1.6%	1.0%	0.7%
8	84.0%	57.4%	35.5%	21.6%	13.5%	8.8%	6.0%	3.1%	1.8%	1.2%
9	95.8%	78.1%	54.7%	35.9%	23.3%	15.4%	10.5%	5.3%	3.1%	2.0%
10	99.2%	91.0%	72.3%	51.7%	35.4%	24.1%	16.7%	8.6%	4.9%	3.1%
11	99.9%	97.0%	85.1%	66.7%	48.7%	34.6%	24.5%	12.9%	7.4%	4.6%
12	*	99.2%	93.0%	78.9%	61.6%	45.8%	33.5%	18.2%	10.5%	6.6%
13	*	99.8%	97.0%	87.7%	72.8%	56.9%	43.1%	24.4%	14.4%	9.0%
14	*	*	98.9%	93.3%	81.8%	67.1%	52.8%	31.3%	18.8%	11.9%
15	*	*	99.6%	96.6%	88.4%	75.8%	61.9%	38.6%	23.8%	15.3%
16	*	*	99.9%	98.4%	92.9%	82.8%	70.1%	46.0%	29.3%	19.1%
17	*	*	*	99.3%	95.9%	88.2%	77.2%	53.3%	35.0%	23.2%
18	*	*	*	99.7%	97.7%	92.2%	83.0%	60.3%	40.9%	27.7%
19	*	*	*	99.9%	98.8%	94.9%	87.6%	66.7%	46.8%	32.3%
20	*	*	*	*	99.4%	96.8%	91.2%	72.5%	52.6%	37.1%
25	*	*	*	*	*	99.8%	98.8%	91.4%	76.9%	60.7%
30	*	*	*	*	*	*	99.9%	97.9%	90.8%	78.9%
35	*	*	*	*	*	*	*	99.6%	96.9%	90.1%
40	*	*	*	*	*	*	*	99.9%	99.1%	95.8%
45	*	*	*	*	*	*	*	*	99.8%	98.4%
50	*	*	*	*	*	*	*	*	99.9%	99.4%

Nota: Analisi di potenza per un test appaiato per rivelare una differenza di espressione di due volte (up oppure down-regolato) nei campioni con una soglia di significatività di 0.0001; ciò dovrebbe dare approssimativamente 1 falso positivo ogni 10.000 geni sull'array. La potenza dipende in modo critico dal coefficiente di variabilità della popolazione. Per esempio, quando la variabilità della popolazione è il 35%, possiamo raggiungere il 95% di potenza con 15 replicati biologici. Se la variabilità della popolazione fosse del 50%, sarebbero necessari 25 replicati biologici per raggiungere una potenza similare.

TABLE 10.2B: Power Analysis for Unpaired Test for 2-Fold Difference with $\alpha = 0.0001$

Group Size	Population Coefficient of Variation									
	20%	25%	30%	35%	40%	45%	50%	60%	70%	80%
3	1.3%	0.7%	0.4%	0.3%	0.2%	0.1%	0.1%	0.1%	0.1%	0.0%
4	7.9%	3.4%	1.7%	1.0%	0.6%	0.4%	0.3%	0.2%	0.1%	0.1%
5	25.5%	10.9%	5.2%	2.8%	1.7%	1.1%	0.8%	0.4%	0.3%	0.2%
6	50.9%	24.7%	12.2%	6.5%	3.7%	2.3%	1.6%	0.8%	0.5%	0.3%
7	74.1%	42.5%	22.5%	12.3%	7.1%	4.4%	2.9%	1.4%	0.8%	0.6%
8	88.8%	60.5%	35.4%	20.2%	11.8%	7.3%	4.8%	2.3%	1.3%	0.9%
9	95.9%	75.4%	49.0%	29.6%	17.9%	11.2%	7.3%	3.5%	2.0%	1.3%
10	98.7%	86.0%	61.9%	40.0%	25.1%	16.0%	10.5%	5.1%	2.8%	1.8%
11	99.6%	92.6%	72.9%	50.4%	33.0%	21.6%	14.4%	7.0%	3.9%	2.4%
12	99.9%	96.4%	81.6%	60.2%	41.3%	27.7%	18.8%	9.3%	5.1%	3.1%
13	*	98.3%	88.0%	69.0%	49.5%	34.3%	23.7%	11.9%	6.6%	4.0%
14	*	99.3%	92.4%	76.5%	57.4%	41.0%	28.9%	14.8%	8.3%	5.0%
15	*	99.7%	95.4%	82.6%	64.7%	47.7%	34.4%	18.1%	10.1%	6.2%
16	*	99.9%	97.3%	87.4%	71.2%	54.3%	40.0%	21.6%	12.3%	7.5%
17	*	*	98.5%	91.1%	76.9%	60.4%	45.6%	25.3%	14.5%	9.0%
18	*	*	99.1%	93.8%	81.7%	66.1%	51.1%	29.2%	17.0%	10.5%
19	*	*	99.5%	95.8%	85.8%	71.3%	56.4%	33.2%	19.7%	12.3%
20	*	*	99.8%	97.2%	89.0%	76.0%	61.4%	37.3%	22.5%	14.1%
25	*	*	*	99.7%	97.5%	91.4%	81.3%	57.3%	37.8%	24.9%
30	*	*	*	*	99.6%	97.4%	92.2%	73.7%	53.4%	37.4%
35	*	*	*	*	99.9%	99.3%	97.1%	85.2%	67.2%	50.2%
40	*	*	*	*	*	99.9%	99.1%	92.3%	78.2%	61.9%
45	*	*	*	*	*	*	99.7%	96.3%	86.2%	72.0%
50	*	*	*	*	*	*	99.9%	98.3%	91.7%	80.1%

***Nota*:** Analisi della potenza per un test disaccoppiato per rivelare la differenza di espressione genica di almeno 2 volte (up o down-regolata) nei campioni con una soglia di significatività di 0.0001; questo dovrebbe dare approssimativamente un falso positivo su 10000 geni sull'array. L'ampiezza del gruppo è il numero di replicati biologici in ciascun gruppo: se questo fosse uno studio clinico, il numero totale dei pazienti dovrebbe essere il doppio dell'ampiezza del gruppo. Questa tavola assume anche che i due gruppi siano di uguale ampiezza (progetto bilanciato). Se i gruppi fossero di ampiezza diseguale, la potenza decrescerebbe. La potenza dipende in modo critico dal coefficiente di variabilità della popolazione. Per esempio, quando la variabilità della popolazione è il 35%, possiamo raggiungere il 95% di potenza per una ampiezza del gruppo di 19. Se la variabilità della popolazione fosse del 50%, potrebbe essere richiesta un'ampiezza del gruppo di 35 per raggiungere potenze similari.

Esempio 10.5 Calcolo della Potenza di uno Studio

Venti pazienti con tumore al seno sono stati trattati nel corso di 16 settimane con chemioterapia basata sul farmaco doxorubicina. I campioni sono stati presi prima e dopo il trattamento, e saranno analizzati per geni up- o down-regolati usando un t-test. Stiamo analizzando 6.500 geni e vogliamo non più di un falso positivo. Il coefficiente di variabilità della popolazione è del 50%. Quale è la potenza dell'analisi per l'identificazione di geni up-regolati almeno due volte? Quale differenza di regolazione possiamo rivelare con una potenza del 95%?

Prima di applicare la formula dell'Equazione 10.2, facciamo qualche calcolo preliminare:

Soglia di significatività. Con l'intento di avere solo un falso positivo, scegliamo una soglia di significatività di 1/6500, che è approssimativamente uguale a 0.00015.

Deviazione Standard. Applichiamo l'Equazione 10.1 con v=0.5 per ottenere la deviazione standard di 0.68 in log in base 2.

Delta. Una regolazione differenziale di almeno due volte corrisponde ad una differenza del rapporto logaritmico di 1 in log in base 2.

Per calcolare la potenza di rivelazione per geni regolati differenzialmente almeno due volte, usiamo la formula:

power.t.test(n=20, delta=1, sd=0.68, sig.level=0.00015, type="one.sample", alternative="two.sided")

ed otteniamo una potenza di 0,94, che è il 94%. Questo significa che applicando l'analisi statistica con una soglia di significatività sufficiente a dare approssimativamente il risultato di un falso positivo, la funzione restituirà il 94% dei geni che sono veramente differenzialmente espressi almeno due volte.

Per sapere quale differenza di regolazione può essere rivelata con il 99% di potenza, usiamo la formula:

power.t.test(n=20, power=0.99, sd=0.68, sig.level=0.00015, type="one.sample", alternative="two.sided")

e troviamo *delta* uguale ad 1.16. Il rapporto di espressione differenziale è uguale a $2^{1.16}$ =2.23, così che possano essere rivelati il 99% dei geni che sono espressi almeno 2.23 volte.

Esempio 10.6: Determinazione del numero di pazienti necessari ad uno studio

In un nuovo studio di chemioterapia del cancro ala mammella, vogliamo identificare i geni che sono up o down-regolati almeno due volte durante il trattamento con doxorubicina. Analizzeremo 10.000 geni e vogliamo al massimo un risultato falso positivo. Il coefficiente di variabilità nella popolazione è del 50%. Vi sono due possibili progetti sperimentali:

- prendere i campioni dallo stesso paziente prima e dopo la terapia, e sviluppare l'analisi accoppiata sul rapporto logaritmico della espressione del gene nei pazienti
- reclutare due gruppi di pazienti, uno da trattare e uno non da trattare, e sviluppare una analisi non accoppiata delle misure di espressione del gene dai pazienti dei due gruppi.

Vogliamo identificare quanti pazienti siano necessari per identificare il 95% dell'espressione differenziale di almeno 2 volte, e quali progetti sperimentali richiedano meno pazienti da reclutare per l'esperimento.

I calcoli preliminari sono simili a quelli dell'esempio 10.5. Il livello di significatività è 1/10000=0.0001; la deviazione standard è 0.68 e il delta é 1.

Per trovare il numero di pazienti necessari al primo progetto sperimentale, usiamo la formula:

power.t.test(power=0.95, delta=1, sd=0.68, sig.level=0.0001, type="one.sample", alternative="two.sided")

ed otteniamo n=21.48950. Un numero frazionario di pazienti non ha alcun significato, quindi approssimiamo questo numero per difetto e concludiamo che sono necessari 22 pazienti per raggiungere la potenza desiderata.

Per trovare il numero di pazienti necessari per il secondo progetto sperimentale, usiamo la formula:

power.t.test(power=0.95, delta=1, sd=0.68, sig.level=0.0001, type="two.sample", alternative="two.sided")

ed otteniamo n=32.15861. Questa è la grandezza del gruppo, e cioè abbiamo bisogno di due gruppi di 33, che equivale a dire un totale di 66 pazienti.

Necessitiamo, quindi, di meno pazienti per l'analisi accoppiata che non per l'analisi non accoppiata, così che il primo progetto sperimentale è meglio del secondo. Tutto ciò illustra un principio generale che le analisi accoppiate sono di solito più potenti delle analisi non accoppiate. Con l'analisi accoppiata, la differenza della espressione del gene viene calcolata con due misure dallo stesso paziente, così che la variabilità individuale è cancellata. Con l'analisi non accoppiata, noi compariamo la media della espressione del gene nei due gruppi; le variabilità tra individui contribuiscono a ciascuna delle medie. A causa di ciò, l'analisi non accoppiata è meno potente dell'analisi accoppiata.

Riassunto dei punti chiave

- Usare la tecnica del blocco per rimuovere il confondimento delle variabili
- Usare la randomizzazione e l'accecamento per rimuovere il problema della polarizzazione
- Evitare campioni di riferimento quando si comparano due campioni dello steso individuo
- Usare i campioni di riferimento per comparare parecchi individui, oppure le serie temporali
- Evitare tecnologie a singolo canale per le serie temporali dove potrebbero esserci cambiamenti globali della espressione del gene
- Calcolare il numero di replicati biologici usando l'analisi della potenza