UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

CORSO DI LAUREA IN STATISTICA

TESI DI LAUREA MAGISTRALE

Integrazione di livelli di espressione e metilazione genica attraverso l'analisi di pathway

Relatore
Prof.ssa Chiara Romualdi
Dipartimento di Biologia

Laureando
Eros Magro
Matricola N. 1038485

Anno Accademico 2013 / 2014

Alla mia Famiglia

Indice

Introduzione

Dalla fine degli anni settanta in cui venivano pubblicate le prime sequenze di acidi nucleici ad oggi l'avanzamento tecnologico ha reso disponibile una sempre maggiore quantità di dati genomici ed epigenetici. Lo studio e l'integrazione dei diversi tipi di dati attraverso tecniche statistiche è un argomento molto attuale e gioca un ruolo fondamentale nel passaggio da un approccio classico della medicina, che propone la stessa cura a tutti i pazienti che hanno la stessa diagnosi, ad un approccio di medicina più personalizzata che invece cerca di tenere in considerazione anche le specifiche caratteristiche genetiche ed epigenetiche specifiche dei singoli individui. Tale approccio alla medicina ha lo scopo di portare a somministrare cure migliori e più sicure. Questo è di fondamentale importanza in particolar modo nelle patologie più complesse ed eterogenee, come ad esempio nel caso del carcinoma ovarico che è trattato in questa tesi tenendo conto dei livelli di espressione e di metilazione dei geni che codificano per delle proteine e dei livelli di espressione dei microRNA.

La tesi è suddivisa in cinque capitoli: nel primo vengono introdotti i concetti biologici utili per la comprensione dei fenomeno che si va ad analizzare, nel secondo vengono esposti i metodi utilizzati nelle analisi, nel terzo si presentano la patologia e i dati trattati, nel quarto si espongono le analisi fatte e i risultati ottenuti mentre l'ultimo è un breve sunto conclusivo.

Capitolo 1

INTRODUZIONE BIOLOGICA

1.1 DAL DNA ALLE PROTEINE

Il DNA, ovvero l'acido desossiribonucleico, è un acido nucleico che contiene le informazioni per la sintesi dell'mRNA (RNA messaggero) e quindi delle proteine. Esso è una macromolecola, o più precisamente un polimero costituito da monomeri detti nucleotidi, ognuno dei quali è composto da un gruppo fosfato, dal desossiribosio (uno zucchero) e da una base azotata. Le quattro basi azotate presenti nel DNA sono l'adenina, la timina, la citosina e la guanina o più sinteticamente A, T, C, G. Queste basi, con una metafora, sono le lettere dell'alfabeto usato per comporre il testo che contiene le informazioni fondamentali per la vita della cellula.

La struttura del DNA è a doppia elica e ognuna di queste è composta da nucleotidi legati con legami covalenti attraverso gruppi fosfato, tale tipo di legame prende il nome di legame fosfodiesterico. Le due eliche sono unite da legami a idrogeno (quindi più deboli) che si formano tra coppie di basi complementari, dunque tra adenina e timina e tra citosina e guanina. La struttura a doppia elica sembra ridondante, dato che da una è possibile ricavare l'altra, ma essa è utile perché rende più stabile la molecola.

Le sequenze del DNA codificanti per una o più proteine sono dette geni,

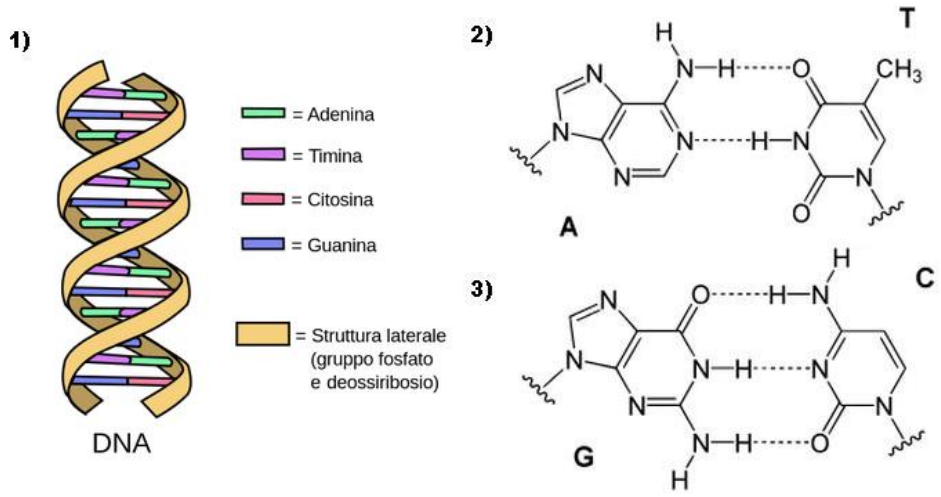

Figura 1.1: In 1) si mostra una rappresentazione semplificata della struttura a doppia elica del DNA. In 2) e 3) si mostrano come i legami a idrogeno leghino rispettivamente l'adenina alla timina e la citosina alla guanina.

mentre con mRNA si intende un filamento simile ad un pezzo di una singola elica di DNA ma con uno zucchero diverso, il ribosio, e con la timina sostituita dall'uracile (U). La fase di copiatura di una sequenza di DNA che porta a una molecola di mRNA viene detta trascrizione. La trascrizione genera una molecola di pre-mRNA identica in sequenza a uno dei due filamenti di DNA, che viene detto filamento codificante o non stampo, mentre il filamento complementare che viene utilizzato per la sintesi è detto filamento stampo. Tale reazione di trascrizione è catalizzata dall'RNA polimerasi, inizia quando quest'enzima si lega ad una specifica regione, detta promotore, che si trova prima del gene e termina quando l'enzima, dopo essersi spostato lungo lo stampo, raggiunge una sequenza terminatrice detta terminatore. La sequenza di pre-mRNA deve poi subire un processo di maturazione che include anche lo splicing, ovvero la rimozione degli introni, che sono pezzi di gene che non vengono tradotti in proteine. Processi di splicing alternativo possono portare a sintetizzare mRNA maturi diversi e quindi proteine diverse da uno stesso gene.

Ottenuto così l'mRNA maturo, ogni gruppo di tre basi, detto tripletta o

codone, viene tradotto in un corrispettivo amminoacido attraverso un insieme di regole detto codice genetico. Ne deriva quindi una catena lineare di amminoacidi che va poi a costituire una struttura tridimensionale detta proteina. È possibile anche che pi strutture tridimensionali si uniscano a formare un'altra proteina che in tal caso si dice avere una struttura quaternaria. Le proteine sono macromolecole essenziali per la vita e svolgono una moltitudine di funzioni biologiche.

1.2 L'ESPRESSIONE GENICA

Il livello di espressione di un gene in una determinata condizione sperimentale è dato dalla quantità di mRNA che produce. Preso come dato assoluto è poco interessante, ma se utilizzato per confrontare due tipi di tessuto differenti, ad esempio con diverse condizioni patologiche, può aiutare a capire l'espressione di quali geni può portare da una condizione all'altra. Con geni differenzialmente espressi (DEG) si intendono quei geni che hanno livelli di espressione significativamente diversi in condizioni biologiche diverse. Fissata una condizione di riferimento i geni differenzialmente espressi possono essere classificati in sovraespressi o sottoespressi.

Tra i geni che possono essere differenzialmente espressi possono esserci degli oncogeni e degli oncosoppressori. Gli oncogeni codificano per delle proteine che aumentano la velocità di replicazione cellulare e gli oncosoppressori codificano delle proteine che la rallentano. Lo studio dei livelli di espressione è di interesse anche per lo studio del cancro in quanto un tumore è una massa di tessuto che cresce in eccesso ed in modo scoordinato rispetto ai tessuti normali.

Il livello di espressione di un gene può dipendere da processi interni alla cellula, in questa tesi se ne tratteranno due molto importanti in ambito epigenetico (ovvero che non modificano la sequenza del DNA): la metilazione e i microRNA (miRNA).

1.2.1 LA METILAZIONE

La metilazione del DNA è un processo che va a modificare l'espressione di un gene a livello pre-trascrizionale. Essa dal punto di vista chimico consiste nell'aggiunta di un gruppo funzionale metile (-CH_3) al quinto carbonio di una cisteina trasformandola in una 5-metilcisteina (Figura 1.2). Questo processo è possibile solamente quando la cisteina fa parte di un sito CpG, ovvero di una sequenza composta da una citosina, un gruppo fosfato e una guanina. Una sequenza di DNA che ha un alto numero di siti CpG viene detta isola CpG. Tali isole si trovano frequentemente nei promotori dei geni dei mammiferi. Un gene con un promotore molto metilato tende ad non essere trascritto. Tale processo è comunque reversibile, ed è utile alla cellula per adattarsi all'ambiente.

Esiste anche un tipo di metilazione che agisce a livello proteico metilando i filamenti degli istoni, le proteine strutturali attorno a cui è avvolto il DNA. Tale metilazione ha l'effetto di aumentare il grado di avvolgimento del DNA impedendo dunque la trascrizione delle regione coinvolta. Anche questo processo è reversibile. In questa tesi non verrà tenuto conto di tale fenomeno.

Si ha dunque che l'ipermetilazione degli oncosoppressori e l'ipometilazione degli oncogeni può portare allo sviluppo di una massa tumorale.

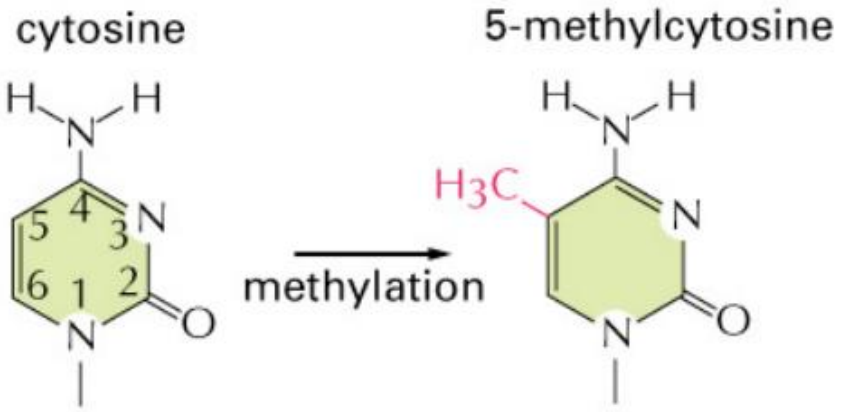

Figura 1.2: Metilazione di una cisteina.

1.2.2 I MICRORNA

I microRNA sono delle corte sequenze di RNA, la cui lunghezza può variare dai 21 ai 25 nucleotidi, che vengono trascritte dal DNA ma che non sono codificati in proteine. Per tale motivo in passato essi non sono stati considerati molto importanti dagli studiosi, in quanto si riteneva erroneamente che l'unico scopo dell'RNA fosse quello di essere tradotto in proteine. Era dunque una prassi non prendere in considerazione i livelli di espressione dei geni che producevano microRNA e se venivano rilevati degli RNA molto corti sconosciuti li si archiviava come frammenti di mRNA più lunghi che erano stati degradati.

Tale convinzione incominciò a essere ridiscussa nel 1993 quando un gruppo di ricerca studiando l'organismo modello C. elegans, un verme lungo circa un 1 mm, fece una scoperta interessante: con una significativa trascrizione del gene *Lin*-4 si assisteva a una sistematica mancanza dell'mRNA e della proteina associati al gene *Lin*-14. Ci si accorse che il gene *Lin*-4, una volta sintetizzato l'RNA maturo, dava luogo a una sequenza di 22 nucleotidi che era complementare ad una specifica regione del gene *Lin*-14. Nonostante lo scalpore, dato che *Lin*-4 era un gene specifico di C. elegans e che quindi non si poteva rilevare tale fenomeno anche in altre specie, tale scoperta non venne presa in debita considerazione. Affinché l'idea che i miRNA giochino un ruolo importante nell'espressione genica prenda piede bisognerà attendere il 2000 con la scoperta del microRNA trascritto dal gene *Let*-7 sempre nell'organismo C. elegans. Esso tuttavia, al contrario di *Let*-4, è un gene presente in molte altre specie, tra cui la Drosophila, l'ape europea, il topo domestico e persino nell'uomo. Tale scoperta è stata da stimolo per la ricerca e lo studio di nuovi geni che venissero trascritti in microRNA.

I microRNA quindi, diversamente da quello che accade con la metilazione, influenzano l'espressione genica a livello post-trascrizionale. Un microRNA va dunque a bloccare la codifica di un mRNA target appaiandosi più o meno perfettamente a una sua determinata sequenza. Se l'appaiamento è perfetto si avrà la degradazione del mRNA target, se invece è imperfetto il target non viene degradato ma solamente inibito. Si ha dunque che un determi-

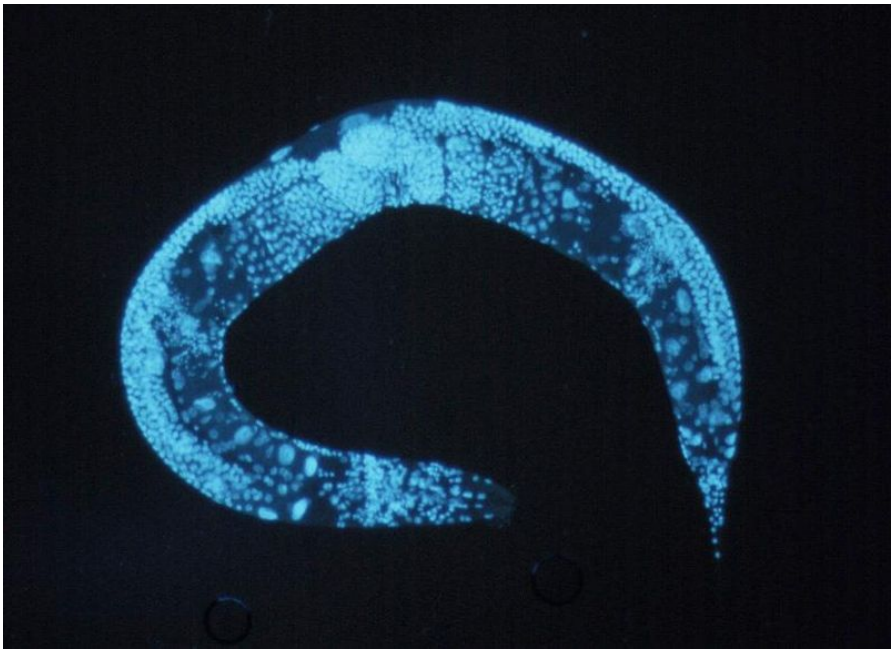

Figura 1.3: Fotografia dell'organismo modello Caenorhabditis elegans.

nato microRNA influenza negativamente l'espressione di un gene, che viene detto target, non permettendo la traduzione dell'mRNA che trascrive. Se i geni target sono degli oncosoppressori aumenta il possibile sviluppo di masse tumorali. Per lo studio di questi ed altri fenomeni biologici è di interesse quindi studiare la differenziale espressione dei microRNA tra diverse condizioni sperimentali, dove il livello di espressione di un microRNA in una data condizione è dato dalla quantità in cui esso viene prodotto.

È importante sottolineare inoltre che un microRNA non ha necessariamente un unico mRNA target e che un mRNA può essere il target di più microRNA, si tratta quindi di una relazione molti a molti che va a formare una fitta rete.

1.3 I MICROARRAY

I microarray sono una tecnologia nata a metà degli anni novanta che può essere utilizzata per la misura dell'espressione genica attraverso il riconosci-

mento degli RNA, indipendentemente dal fatto che questi vengano tradotti in proteine. Sono quindi una importante risorsa per raccogliere i dati relativi ai livelli di espressione dei geni codificanti e dei microRNA. Si tratta di una tecnologia che sfrutta una tecnica di ibridazione inversa per monitorare l'espressione genica di decine di migliaia di geni con un unico esperimento.

Un microarray è costituito da un insieme di tante piccole sonde costituite da frammenti di DNA a singola elica, detti probe, fissati ad una superficie solida che può essere di vetro, di plastica o un chip di silicio.

Nel caso dei geni codificanti ogni tipo di probe corrisponde ad un unico gene e la sequenza di un probe è data dal complementare dalla parte iniziale della sequenza del gene che gli corrisponde. Come già spiegato nel paragrafo 1.1 un gene codificante può produrre diversi tipi di mRNA grazie al fenomeno degli splicing alternativi e questo può portare a pensare che in realtà un probe non possa sempre rappresentare tutti i tipi di mRNA prodotti da un gene codificante. In realtà non è così in quanto le sequenze iniziali di tutti i tipi di mRNA che possono essere prodotti da un gene sono uguali tra loro, si ha quindi che ad un probe che rappresenta un gene codificante corrispondono tutti i possibili tipi di mRNA che questo può produrre essendo la sua sequenza identica alle loro parti iniziali. Nel caso dei microRNA invece la sequenza di un probe è identica sia in lunghezza che in sequenza a quella del microRNA a cui è associato.

I tipi di probe da fissare al supporto sono scelti a priori a partire, per esempio, da delle librerie biologiche. L'insieme di probe uguali adiacenti è detto probeset o spot. Tali probset vanno quindi a formare una matrice sulla superficie del microarray: ad ogni probset corrisponde un gene o un miRNA.

Va ora fatta una distinzione tra tecnologia a singolo canale e tecnologia a doppio canale. La tecnologia a singolo canale permette di rilevare un livello di espressione assoluto per ogni probset in una determinata condizione, come può essere ad esempio quella di un tessuto malato, mentre la tecnologia a doppio canale rileva l'espressione relativa a due condizioni diverse, dando quindi per ogni probeset il rapporto dei livelli di espressione tra le due, come ad esempio il rapporto tra i livelli di espressione di un tessuto sano e i livelli

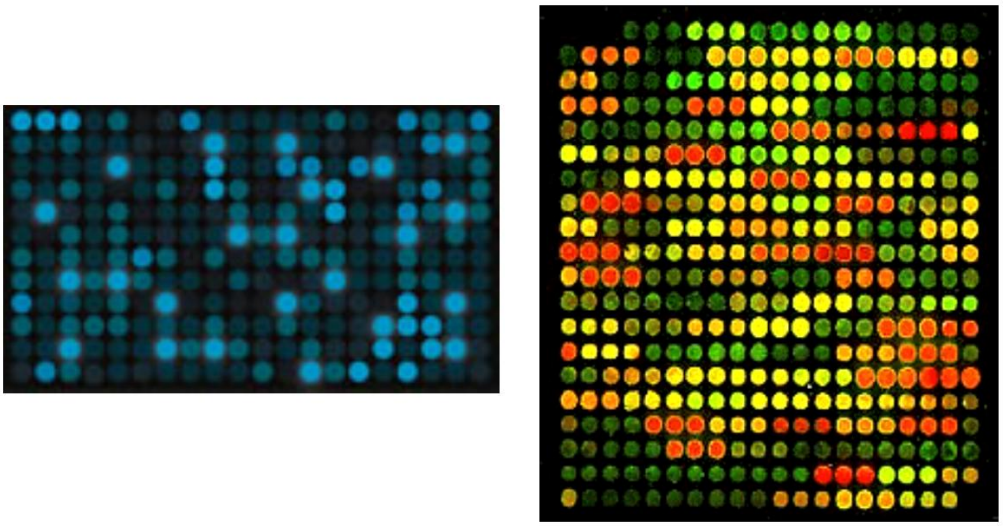

Figura 1.4: Immagini relative a due esperimenti microarray. Quella di sinistra è relativa a un microarray a singolo canale e rappresenta dei livelli di espressione genica assoluti, quella di desta è relativa a un microarray a doppio canale e rappresenta dei livelli di espressione genica relativi. Nella prima si notano diverse intensità luminose, nella seconda si nota che gli spot vanno dal verde al rosso passando per il giallo, che rappresenta un eguale livello di espressione in entrambe le condizioni.

di espressione di un tessuto malato.

Si spiega ora come viene effettuata tale rilevazione per un singolo esperimento. Innanzitutto la prima cosa da fare è estrarre il trascrittoma (mRNA, miRNA) dal o dai tessuti di interesse a seconda che si lavori con il singolo o con il doppio canale. Una volta che è stato ottenuto questo viene convertito in cDNA (DNA complementare) grazie all'aiuto di un enzima detto trascrittasi inversa. Il cDNA viene poi marcato con una sonda fluorescente nel caso del singolo canale o con sonde fluorescenti di diverso colore (rosso e verde) nel caso del doppio canale, dove i colori diversi rappresentano il tessuto di provenienza. Fatto questo, quello che si è ottenuto viene posto sul microarray in modo tale che possa avvenire l'ibridazione tra le sonde e i cDNA secondo la regola delle basi complementari. Una volta che l'ibridazione ha avuto luogo il microarray viene lavato e posto all'interno di un apposito scanner ottico che,

eccitando i fluorofori utilizzati per marcare il cDNA, attraverso un sistema informatico va a ottenere una immagine ad alta definizione per ogni probset. Da tali immagini bisogna poi ricavare un valore che rappresenti un livello di espressione. Per fare questo per ognuna di esse viene rilavata la gradazione del colore di ognuno dei pixel che le compongono al fine di costruire delle curve delle distribuzioni di ogni segnale. Si può poi prendere la mediana o la media di ogni distribuzione per avere un singolo livello di espressione per ogni spot.

Dal punto di vista economico la tecnologia a doppio canale risulta più economica perché a parità di numero di esperimenti da effettuare utilizza la metà dei microarray. Essa inoltre permette anche, effettuando un paragone tra due condizioni nello stesso esperimento, un maggior controllo dell'effetto di fattori sistematici esterni non biologici che vanno a influenzare il risultato; d'altro canto però va tenuto conto che il fluoroforo verde e il fluoroforo rosso non hanno la stessa efficienza. Inoltre il singolo canale permette di ottenere risultati più riproducibili e di confrontare anche più condizioni in modo semplice. Infine gli effetti sistematici non biologici possono essere risolti attraverso la normalizzazione dei dati, come verrà spiegato in seguito. Per questi motivi la tecnologia a singolo canale sta prendendo sempre più piede rispetto a quella a doppio canale.

1.4 LE MATRICI DEI DATI

In questo paragrafo si definisce la struttura delle metrici dei dati biologici che sono utilizzate in questa tesi. I dati a disposizione sono relativi a n pazienti comuni, di cui n_1 sono nella condizione biologica 1 e n_2 sono nella condizione biologica 2.

Per il livello di espressione dei geni codificanti si definisce la matrice

$$Y = \left[\begin{array}{ccc|ccc} y_{1,1} & \cdots & y_{1,n_1} & y_{1,n_1+1} & \cdots & y_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{G,1} & \cdots & y_{G,n_1} & y_{G,n_1+1} & \cdots & y_{G,n} \end{array}\right] = \left[\begin{array}{c|c} Y_1 & Y_2 \end{array}\right] \qquad (1.1)$$

dove ogni riga rappresenta un gene, ogni colonna rappresenta una paziente, Y_1 e Y_2 rappresentano rispettivamente gli esperimenti relativi alla condizione biologica 1 e 2, e ogni y_{ij} è il logaritmo del livello di espressione assoluto rilevato da un esperimento microarray a singolo canale. Il logaritmo è giustificato dal fatto che il range di valori che può restituire un microarray va da 0 a 2^{16} e dunque, essendo una funzione monotona che cresce lentamente, stabilizza la varianza pesando meno le differenze tra valori grandi rispetto alle medesime differenze tra valori piccoli.

Per il livello di espressione dei microRNA si definisce la matrice

$$Y^{(m)} = \left[\begin{array}{ccc|ccc} y^{(m)}_{1,1} & \cdots & y^{(m)}_{1,n_1} & y^{(m)}_{1,n_1+1} & \cdots & y^{(m)}_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y^{(m)}_{G_m,1} & \cdots & y^{(m)}_{G_m,n_1} & y^{(m)}_{G_m,n_1+1} & \cdots & y^{(m)}_{G_m,n} \end{array}\right] = \left[\begin{array}{c|c} Y^{(m)}_1 & Y^{(m)}_2 \end{array}\right] \tag{1.2}$$

dove ogni riga rappresenta un microRNA, ogni colonna rappresenta una paziente, $Y^{(m)}_1$ e $Y^{(m)}_2$ rappresentano rispettivamente gli esperimenti relativi alla condizione biologica 1 e 2, e ogni $y^{(m)}_{ij}$ è il logaritmo del livello di espressione assoluto rilevato da un esperimento microarray a singolo canale per le stesse considerazioni di prima.

L'analisi del grado di metilazione dei promotori dei geni codificanti può essere svolto con un procedimento che sfrutta l'ibridazione inversa simile a quello per l'RNA. Essa produce per ogni gene analizzato una misura di intensità di metilazione (ϕ^m) e una misura di intensità di non metilazione (ϕ^{nm}). La quantità $\phi^m/(\phi^m + \phi^{nm})$ si dice *β-value* e rappresenta la probabilità che un gene codificante non sia trascritto a causa della metilazione a livello del DNA. Si definisce dunque per la metilazione la matrice

$$Y^{(M)} = \left[\begin{array}{ccc|ccc} y^{(M)}_{1,1} & \cdots & y^{(M)}_{1,n_1} & y^{(M)}_{1,n_1+1} & \cdots & y^{(M)}_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y^{(M)}_{G_M,1} & \cdots & y^{(M)}_{G_M,n_1} & y^{(M)}_{G_M+1,n_1} & \cdots & y^{(M)}_{G_M,n} \end{array}\right] = \left[\begin{array}{c|c} Y^{(M)}_1 & Y^{(M)}_2 \end{array}\right] \tag{1.3}$$

dove ogni riga rappresenta un gene codificante, ogni colonna rappresenta una paziente, $Y^{(M)}_1$ e $Y^{(M)}_2$ rappresentano rispettivamente gli esperimenti relativi

alla condizione biologica 1 e 2, e ogni $y_{ij}^{(M)}$ è il logit del β-*value* del gene i relativo alla paziente j. Si ha quindi che i valori della matrice sono continui.

Capitolo 2

METODI

2.1 COMPARABILITÀ DEI LIVELLI DI ESPRESSIONE E NORMALIZZAZIONE QUANTILE

Per rendere paragonabili tra loro gli n esperimenti a singolo canale relativi alle matrici di dati dei geni codificanti proteine e dei microRNA è necessario che le matrici vengano normalizzate. La normalizzazione serve a correggere gli effetti di errori sistematici dovuti ad esempio alla diversa potenza del laser dello scanner ottico oppure ad altri parametri di scansione. Tale operazione va quindi effettuata quindi prima di fare qualsiasi tipo di analisi su di esse.

Esistono varie tecniche di normalizzazione ma studi di simulazione, come ad esempio Chiogna *et al.* (2009), hanno mostrato che la normalizzazione adottata influenza in maniera minima i risultati delle analisi. Per questo motivo di seguito si presenterà la normalizzazione quantile, una tra le più comuni e semplici che ha inoltre anche il vantaggio di essere computazionalmente parsimoniosa.

La normalizzazione quantile ha l'obbiettivo di rendere uguali le distribuzioni empiriche degli array di ogni esperimento. Per fare questo si parte dalla considerazione che, se la distribuzione di due vettori di dati è la stessa, il loro grafico quantile-quantile è formato da punti che giacciono perfettamente

sulla diagonale. Estendendo questo concetto a n dimensioni, e quindi ha n esperimenti, si ha che gli n vettori hanno la medesima distribuzione se e solo se il loro grafico quantile-quantile n-dimensionale ha i punti esattamente sulla diagonale dell'ipercubo n-dimensionale. Basta quindi proiettare i punti del grafico quantile-quantile osservato sulla diagonale di questo ipercubo. Per fare questo è sufficiente sostituire i valori della matrice di dati con la media dei valori relativi ai quantili ordinati.

Si spiega ora nel dettaglio l'algoritmo per ottenere una matrice di dati normalizzata. Data una generica matrice dei dati di espressione X di dimensione $p \times n$, dove p è il numero di geni ed n è il numero di esperimenti, si proceda nel seguente modo:

1. si ordinino gli elementi di ogni colonna di X in modo da ottenere X^{ord};
2. si calcolino ora le medie delle righe di X^{ord} ottenendo il vettore $p \times 1$ $\overline{x}$;
3. si costruisca la matrice $p \times n$ $X^*_{ord} = [\overline{x}, ..., \overline{x}]$;
4. si riordinino gli elementi delle colonne di X^*_{ord} in modo tale che abbiano lo stesso ordine che avevano nella matrice X, ottenendo così la matrice normalizzata.

Da questo punto in poi si assumerà che le matrici (1.1) e (1.2) siano state normalizzate.

2.2 UN APPROCCIO UNIVARIATO

Quello che è di interesse fare in questo paragrafo è di esporre un metodo per l'identificazione dei geni codificanti proteine che sono differenzialmente espressi. Si è anche interessati all'identificazione dei geni differenzialmente metilati e dei microRNA differenzialmente espressi, ma il problema da risolvere dal punto di vista statistico è lo stesso.

Il primo e più semplice metodo che viene in mente è quello di paragonare la popolazione nella condizione 1 e quella nella condizione 2 con il classico test t

per verificare l'uguaglianza in media. Tale test andrebbe quindi eseguito su ogni gene, prendendo ad esempio i geni codificanti.

Un primo problema che ha tale approccio è che i valori di espressione variano da valori prossimi allo zero a valori molto grandi. Non è quindi infrequente che geni con valori di espressione bassi abbiano anche valori di varianza molto bassi e di conseguenza valori molto alti del test t che portano a rifiutare l'ipotesi nulla anche se questa è vera. Per risolvere tale problema si utilizzerà una statistica test t moderata, ovvero il test bayesiano empirico (test Ebayes).

Un ulteriore problema è dato dal fatto di controllare l'errore globale visto che il numero di test che si vanno a svolgere è generalmente molto alto, per risolvere tale problema si propone una tecnica basata sul controllo dell'FDR.

Per semplicità di notazione in questo paragrafo si esporranno i metodi facendo riferimento alla matrice di dati (1.1).

2.2.1 TEST EBAYES

Dati gli n esperimenti della matrice Y si definisce il vettore della variabile risposta per ogni gene g come $y_g^T = (y_{g1}, ..., y_{gn})$. Per ogni y_g si assume un modello lineare del tipo

$$y_g = X\alpha_g + \epsilon_g$$

dove X è una matrice di disegno a rango pieno, α_g è il vettore dei parametri e ϵ_g è in termine di errore di media zero. Si ha dunque che

$$E(y_g) = X\alpha_g.$$

Si assume inoltre che

$$Var(y_g) = W_g\sigma_g^2$$

dove W_g è una matrice nota definita non negativa di pesi. Tale modello ammette la presenza di valori mancanti, in tal caso la diagonale di W_g può contenere pesi pari a zero per tali valori. Va inoltre precisato che il termine

d'errore ϵ_g, e di conseguenza la variabile risposta y_g, non è assunto normale e che il modello non è necessariamente stimato col metodo dei minimi quadrati.

Si definisce dunque

$$\beta_g = C^T \alpha_g$$

per rappresentare dei contrasti che si assumono essere di interesse biologico. In particolare si assume di essere interessati a testare l'ipotesi $\beta_{gj} = 0$.

Il modello lineare per ogni gene porge lo stimatore dei parametri $\hat{\alpha}_g$, lo stimatore s_g^2 di σ_g^2 e la matrice di covarianza

$$Var(\hat{\alpha}_g) = V_g s_g^2$$

dove V_g è una matrice definita positiva non dipendente da s_g^2. Lo stimatore dei contrasti è dunque

$$\hat{\beta}_g = C^T \hat{\alpha}_g$$

con matrice di covarianza

$$Var(\hat{\beta}_g) = C^T V_g C s_g^2.$$

Assumendo ora che

$$\hat{\beta}_{gj} | \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$$

e

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi^2_{d_g}$$

dove v_{gj} è il j-esimo elemento diagonale di $C^T V_g C$ e d_g sono i gradi di libertà del modello lineare per il gene g, si ha che il test t ordinario

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{v_{gj}}}$$

sotto l'ipotesi nulla approssimativamente si distribuisce come una t si Student con d_g gradi di libertà.

Fin qui l'analisi non ha tenuto conto della struttura parallela dei test che si vanno ad eseguire. Per fare questo si descrive come i coefficienti β_{gj}

e le varianze σ_g^2 variano attraverso i geni. Si costruisce dunque un modello gerarchico che assume delle distribuzioni a priori che ora verranno specificate.

La distribuzione a priori assunta su σ_g^2 è

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2$$

e descrive come varia la varianza tra i geni. Per ogni dato j si assume inoltre che β_{gj} sia diverso da zero con probabilità nota

$$P(\beta_{gj} \neq 0) = p_j,$$

dove p_j rappresenta la percentuale che ci si aspetta di geni differenzialmente espressi. Si assume infine che

$$\beta_{gj} | \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$$

per descrivere la distribuzione dei $fold\ changes$ per i geni che sono differenzialmente espressi.

Utilizzando tale approccio gerarchico si ha che la media a posteriori di $\sigma_g^{-2} | s_g^2$ è

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}.$$

Ne deriva che la statistica test t moderata che si ottiene è

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}.$$

Tale statistica sotto l'ipotesi nulla $H_0 : \beta_{gj} = 0$ si distribuisce come una t di Student con $d_g + d_0$ gradi di libertà. I gradi di libertà aggiunti a $\tilde{t}_{gj}$ rispetto a t_{gj} rispecchiano le informazioni extra che sono state inserite nel modello. Si noti inoltre che la statistica test t moderata si riduce alla statistica test t ordinaria quando $d_0 = 0$.

Resta da scegliere il valore dei parametri s_0 e d_0. Questi vengono stimati dai dati, ciò motiva il fatto che questo test bayesiano venga detto empirico. La

stima avviene con il metodo dei momenti, eguagliando i primi due momenti teorici della variabile $\log(s_g^2)$ con quelli empirici. Per maggiori dettagli si veda (Smyth, 2004).

2.2.2 FDR

Quando si effettuano molti test statistici è noto che la significatività α di ogni singolo test, ovvero la probabilità di rifiutare l'ipotesi nulla quando questa è vera, non è pari alla significatività del test globale. Se si effettuano G verifiche d'ipotesi, ognuna con livello di significatività pari ad α, con test tra loro indipendenti, se tutte le ipotesi sono nulle la probabilità di commettere almeno un errore è $1 - (1 - \alpha)^G$. Si ha quindi che tale quantità è il livello di significatività del test globale, detto anche livello di copertura totale. Se invece i singoli test sono dipendenti si ha che il livello di copertura può essere sia più grande che più piccolo del livello che si avrebbe nel caso in cui i test singoli fossero dipendenti. L'unica cosa che si può affermare in tal caso è che esso è minore o uguale alla somma dei livelli di significatività dei singoli test.

Dato che in ambito genetico si testano anche decine di migliaia di geni è necessario adottare delle tecniche per tenere sotto l'errore globale. La maggior parte di queste per la correzione dei livelli di significatività, come ad esempio la correzione di Bonferroni, la correzione di Holm e la correzione di Holm-Sidak, si basano sul controllo del *Family Wise Error Rate* (FWER). Si definisce l'FWER come la probabilità di avere almeno un falso positivo tra i test effettuati, ovvero, con riferimento alla Tabella 2.1, $FWER = P(V \geq 1)$. Quando il numero di test che si esegue è molto alto, come nel caso che si sta trattando, si tratta di una misura molto conservativa.

	Accetto H_0 $(-)$	Rifiuto H_0 $(+)$	Totale
H_0 vera (-)	U	V	G_0
H_0 falsa $(+)$	T	S	$G - G_0$
Totale	$G - R$	R	G

Tabella 2.1: Matrice di confusione che rappresenta i possibili esiti di G test statistici.

In questa tesi si tratterà un approccio diverso che risulta essere migliore, basato sul controllo di una quantità diversa detta False Discovery Rate (FDR). L'FDR la è frazione attesa di falsi positivi nella lista dei test che rifiutano l'ipotesi nulla, ovvero, per esempio, la percentuale di geni che nella realtà non sono differenzialmente espressi all'interno della lista di geni differenzialmente espressi individuata. Utilizzando la notazione della Tabella 2.1 si ha che

$$FDR = E\left[\frac{V}{R}|R > 0\right] P(R > 0). \tag{2.1}$$

La misura di significatività basata sull'FDR è il q-value. Per chiarire la differenza tra p-value e q-value si propone il seguente esempio. Si supponga di avere due liste di geni differenzialmente espressi, una ricavata utilizzando un p-value del 5%, l'altra utilizzando un q-value del 5%. Nel primo caso in media il 5% dei geni che nella realtà non sono differenzialmente espressi faranno parte della lista, nel secondo caso invece in media si ha che il 5% dei geni della lista è costituita da geni che nella realtà non sono differenzialmente espressi.

Essendo il numero di test eseguiti elevato si ha che $P(R > 0)$ è un valore prossimo all'unità, si può dunque approssimare la (2.1) nel seguente modo:

$$FDR \doteq E\left[\frac{V}{R}|R > 0\right] \doteq \frac{E[V]}{E[R]}. \tag{2.2}$$

Riscrivendo le quantità V ed R in funzione della soglia t si ha

$$V(t) = \#\{p_i \leq t; i = 1, ..., G; \quad H_0 \; vera\},$$

$$R(t) = \#\{p_i \leq t; i = 1, ..., G\}$$

dove p_i è il p-value dell'i-esimo test. Sostituendo alla (2.2) si ottiene

$$FDR(t) = \frac{E[V(t)]}{E[R(t)]}$$

con $R(t)$ noto e $V(t)$ ignoto e quindi da stimare.

Sfruttando il fatto che la distribuzione dei p-value sotto l'ipotesi nulla è

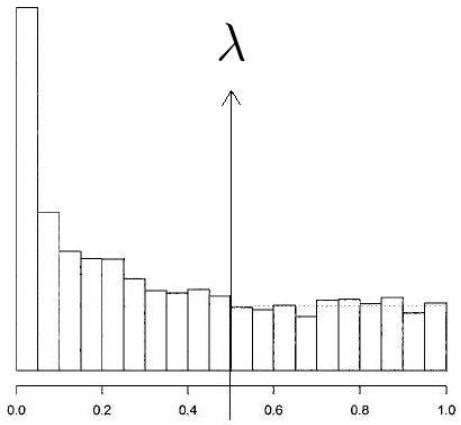

Figura 2.1: Esempio di un possibile istogramma dei p-value dei singoli test in cui si mette in evidenza una plausibile soglia $\lambda = 0.5$ oltre la quale la distribuzione dei p-value è uniforme.

uniforme si può scrivere

$$E[V(t)] = G_0 t.$$

Il problema ora è la stima di G_0 che equivale a stimare $\pi_0 = G_0/G$.

Se si definisce $\hat{\pi}_0$ la stima di π_0 si ha che la stima dell'FDR risulta essere

$$\hat{FDR}(t) = \frac{\hat{\pi}_0 G t}{R(t)}.$$

Essendo il q-value definito come il minimo FDR che si ottiene quando si definisce un test significativo è possibile scriverne la stima come

$$\hat{q}(p_i) = \min_{p_i \leq t} \hat{FDR}(t), \quad i = 1, ..., G. \tag{2.3}$$

Se si ordinano i p-value in modo crescente si può riscrivere la (2.3) nella seguente forma ricorsiva indipendente dalla soglia t utilizzata

$$\begin{cases} \hat{q}(p_{(i)}) = \min\left(\frac{(\hat{\pi}_0 G p_{(i)})}{i}, \hat{q}(p_{(i+1)})\right) & i = 1, ..., G-1 \\ \hat{q}(p_{(i)}) = \hat{\pi}_0 p_{(i)} & i = G \end{cases}.$$

Resta dunque da stimare $\hat{\pi}_0$. Il primo approccio utilizzato da Benjami-

ni e Hochberg (1995) fu quello di considerare la situazione più conservativa possibile ponendo $\hat{\pi}_0 = 1$. Successivamente Storey e Tibshirani (2001) hanno proposto di cercare nella distribuzione dei p-value un valore soglia λ oltre il quale la distribuzione è uniforme (si veda l'esempio della Figura 2.1) e di porre la stima di $\hat{\pi}_0$ pari alla proporzione di p-value che superano tale soglia.

2.3 MODELLI GRAFICI GAUSSIANI

Diversamente dal paragrafo precedente dove si è proposto un approccio univariato per ogni singolo gene qui si propone un approccio di *gene-set analysis* (per una review sui metodi di *gene-set analysis* si veda Khatri *et al.*, 2012), ovvero che si basa su dei test multivariati effettuati su un insieme di geni predefinito. Nel caso specifico questo insieme di geni è relativo a una via metabolica o di segnale, detta *pathway*.

Un *pathway* è un grafo che modella un determinato processo biologico in cui i nodi rappresentano proteine, geni, metaboliti (ulteriori specie chimiche intermedie o finali del processo in questione) ed altre possibili entità. Le relazioni che vi possono essere tra le entità rappresentate dai nodi all'interno del processo biologico possono essere di varia natura (interazioni, inibizioni, attivazioni, fosforilazioni, trasformazioni...) e sono rappresentate con degli archi. Esistono basi di dati biologiche che contengono liste di *pathway*, come ad esempio la *Kyoto Encyclopedia of Genes and Genomes* (KEGG).

Questo paragrafo ha lo scopo di introdurre un metodo di *gene-set analysis* basato sui modelli grafici gaussiani, proposto da Massa *et al.* (2010) e Martini *et al.* (2013), per trattare i livelli di espressione (o eventualmente di metilazione) dei geni codificanti proteine includendo le informazioni topologiche date a priori da uno specifico *pathway*. Nella prima parte si esporranno degli utili concetti introduttivi sui grafi, nella seconda si spiegherà come utilizzare i modelli grafici gaussiani per fare dei test sul *pathway* a livello globale mentre nella terza si tratterà un metodo per identificare quali parti del *pathway* sono maggiormente coinvolte nel passaggio da una condizione biologica ad un'altra.

Per semplicità di notazione nella seconda e nella terza parte si esporrà il metodo facendo riferimento alla matrice (1.1).

2.3.1 GRAFI

Un grafo $\mathcal{G}$ è definito dalla coppia (V, E) dove V è un insieme finito di nodi (vertici) ed E è un insieme di archi, più precisamente si ha che $E \subseteq V \times V$. Un generico arco $(u, v) \in E$ si dice non orientato se si ha che $(v, u) \in E$, viceversa se $(v, u) \notin E$ si dice che l'arco è orientato. Se un grafo ha solo archi non orientati si dice non orientato, se invece ha solo archi orientati si dice orientato. In un grafo non orientato se tra i nodi u e v c'è un arco tali nodi sono detti adiacenti. In un grafo orientato invece se $u \to v$, ovvero $(u, v) \in E$, si dice che u è genitore di v e che v è figlio di u.

Si definisce cammino (path) una sequenza di nodi $(v_1, ..., v_k)$ se per ogni $i = 1, ..., k-1$ si ha che $(v_i, v_{i+1}) \in E$. Se in un cammino v_1 e v_k coincidono tale cammino viene detto ciclo. Un grafo orientato privo di cicli o aciclico viene detto DAG. Dato un DAG D, si definisce un grafo morale D^m come un grafo non orientato ottenuto aggiungendo archi non orientati tra tutti i nodi di D che hanno un figlio in comune (se non sono già connessi da un arco) e rendendo tutti gli archi di D non orientati.

Un grafo non orientato si dice completo se per ogni $u \in V$ e per ogni $v \in V$ si ha che v è adiacente ad u, ovvero se ogni nodo del grafo ha un arco in comune con tutti gli altri nodi del medesimo. Un grafo $\mathcal{G}_A = (A, E_A)$ si dice sottografo di $\mathcal{G}$ se $A \subseteq V$ e $E_A = E \cap (A \times A)$. Un sottografo completo che non è contenuto in nessun altro sottografo completo è detto cricca (*clique*).

Se $\mathcal{G}$ è un grafo non orientato la tripla (A, B, C) di sottoinsiemi disgiunti di V si dice decomposizione di $\mathcal{G}$ se $V = A \cup B \cup C$, C è un sottoinsieme completo di V e C separa A e B. Si ha che $\mathcal{G}$ si dice decomponibile se è completo oppure se possiede una decomposizione (A, B, C) tale che entrambi i sottografi $\mathcal{G}_{A \cup B}$ e $\mathcal{G}_{B \cup C}$ siano decomponibili.

Un grafo triangolato è un grafo non orientato in cui ogni ciclo composto da almeno quattro nodi ha due nodi non consecutivi che sono adiacenti. Si può dimostrare che un grafo non orientato è decomponibile se e solo se è

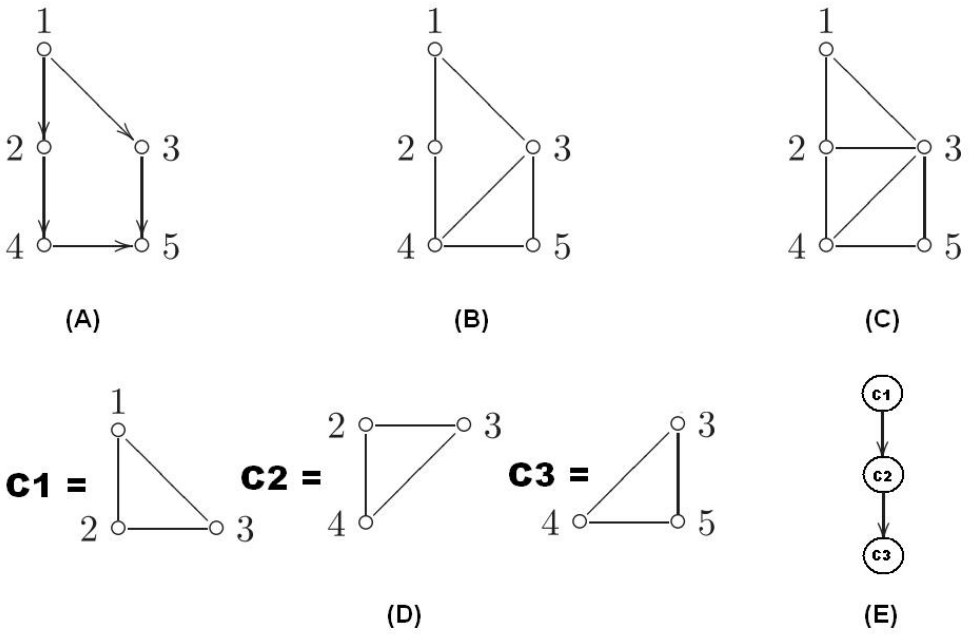

Figura 2.2: In A) si mostra un esempio di DAG, in B) il suo corrispondente grafo morale, in C) una triangolarizzazione del grafo morale, in D) la scomposizione in *clique* del grafo triangolato ed infine in E) il relativo *junction tree*.

triangolato (cfr. Lauritzen, 1996). Se un grafo non è triangolato è possibile renderlo tale aggiungendo degli archi extra. Dal punto di vista informatico ottenere la triangolarizzazione ottima, ovvero con il più piccolo numero di archi da aggiungere, è un problema NP-hard, nella pratica quindi si ricorre a degli algoritmi euristici.

Un *junction tree* per un grafo $\mathcal{G}$ è un albero avente le *clique* di $\mathcal{G}$ come nodi che soddisfa la *running intersection property*, vale a dire che per ogni coppia di *clique* C_i e C_j dell'albero ogni *clique* del cammino che connette C_i e C_j deve contenere $C_i \cap C_j$. La decomponibilità è una condizione necessaria e sufficiente per l'esistenza del *junction tree*.

2.3.2 TEST SULL'INTERO PATHWAY

Dato un *pathway* è necessario che questo sia convertito in un grafo contenente solo nodi che rappresentano geni. Si sostituiscono quindi i nodi delle proteine con i nodi relativi ai geni che le hanno originate e si eliminano i metaboliti. Per non perdere l'informazione dei metaboliti se un nodo era raggiungibile da un altro nodo passando per un metabolita, allora tra i due nodi viene creato un arco. Questo però non viene fatto per tutti i metaboliti perché ve ne sono alcuni di molto frequenti (H^+, H_2O, ATP, ...) che non avrebbe senso considerare.

Ottenuto quindi un grafo di soli geni il passo successivo consiste nel trasformarlo in un DAG e moralizzarlo, ottenendo dunque il grafo finale $\mathcal{G}$. Date le due condizioni biologiche da paragonare sul grafo $\mathcal{G}$ si definiscono i due seguenti modelli grafici gaussiani

$$\mathcal{M}_1(\mathcal{G}) = \{Y_1 \sim N_G(\mu_1, \Sigma_1), K_1 = \Sigma_1^{-1} \in S^+(\mathcal{G})\}$$

$$\mathcal{M}_2(\mathcal{G}) = \{Y_1 \sim N_G(\mu_1, \Sigma_2), K_2 = \Sigma_2^{-1} \in S^+(\mathcal{G})\}$$

dove G è il numero di vertici del grafo (e quindi il numero di geni), $S^+(\mathcal{G})$ è l'insieme delle matrici simmetriche definite positive con elementi nulli in corrispondenza degli archi mancanti in $\mathcal{G}$ e μ_1, μ_2, Σ_1 e Σ_2 sono parametri ignoti.

Definito il modello è di interesse capire se i geni relativi al *pathway* variano l'intensità delle relazioni con gli altri geni nel medesimo al cambiare della condizione biologica. Le informazioni per testare tale ipotesi sono contenute nelle matrici di covarianza del modello o equivalentemente nelle matrici di concentrazione K_1 e K_2. L'ipotesi che si intende testare è dunque

$$\begin{cases} H_0 : K_1 = K_2 \\ H_1 : K_1 \neq K_2 \end{cases} . \tag{2.4}$$

Senza perdita di generalità supponiamo di avere $\gamma_1 = (\gamma_1^j)$, con $j = 1, ..., n_1$ osservazioni da una $N_G(0, \Sigma_1)$ e $\gamma_2 = (\gamma_2^j)$, con $j = 1, ..., n_2$ osservazioni da una $N_G(0, \Sigma_2)$, dove $K_1 = \Sigma_1^{-1} \in S^+(\mathcal{G})$ e $K_2 = \Sigma_2^{-1} \in S^+(\mathcal{G})$.

Definito

$$W_i = \sum_{j=1}^{n_i} (\gamma_i^j)(\gamma_i^j)^T, \quad i = 1, 2$$

la funzione di massima verosimiglianza è data da

$$L(K_1, K_2) = \prod_{i=1}^{2} (2\pi)^{\frac{n_i p}{2}} \det(K_i)^{\frac{n_i}{2}} \exp\{-\frac{1}{2} Tr(K_i W_i)\}.$$

Sotto l'ipotesi alternativa si ha che le stime di Σ_1 e Σ_2 sono date rispettivamente da $S_1 = (n_1 - 1)^{-1} W_1$ e $S_2 = (n_2 - 1)^{-1} W_2$, mentre sotto l'ipotesi nulla si ha che la stima di Σ è data da $S = (n_1 + n_2 - 2)^{-1}\{(n_1 - 1)S_1 + (n_2 - 1)S_2\}$.

Queste stime tuttavia non sono molto affidabili in quanto, come è solito in ambito genetico, il numero di geni G è molto più grande del numero di esperimenti n. Un semplice approccio non parametrico per risolvere il problema consiste nel ridurre la varianza degli stimatori tramite boostrap, ricampionando B volte le colonne dei dati ottenendo quindi B stime e facendo infine la media di queste. Tale approccio comunque quando il numero di geni è molto alto è computazionalmente molto pesante. Per tale motivo, in questo contesto usualmente si usa un metodo di shrinking per le tre matrici da stimare. Con riferimento alla stima S di Σ si ottiene $S^* = [s_{ij}^*]$ dove

$$s_{ij}^* = \begin{cases} s_{ii} & i = j \\ r_{ij}^* & i \neq j \end{cases}$$

con

$$r_{ij}^* = r_{ij} \min(1, \max(0, 1 - \hat{\lambda}^*))$$

e con

$$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{Var}(r_{ij})}{\sum_{i \neq j} r_{ij}^2}$$

dove r_{ij} e s_{ii} rappresentano rispettivamente la correlazione e la varianza empirica. Per maggiori dettagli si veda Schäfer e Strimmer (2005).

Resta ancora un ulteriore problema da risolvere, ovvero vincolare le stime a essere zero in corrispondenza degli archi mancanti. Tale problema viene risolto dall'algoritmo IPS (*Interative Proportional Scaling*) che prende in

input le stime precedentemente ricavate ed impone tale vincolo restituendo dunque le stime finali $\hat{\Sigma}$, $\hat{\Sigma}_1$ e $\hat{\Sigma}_2$. Per approfondimenti di veda Fienberg (1970) e i suoi riferimenti.

Si ha dunque che il test rapporto di verosimiglianza è

$$\Lambda = \frac{L_{H_0}(\hat{K}, \hat{K})}{L_{H_1}(\hat{K}_1, \hat{K}_2)}$$

dove $\hat{K}_1 = \hat{\Sigma}_1^{-1}$, $\hat{K}_2 = \hat{\Sigma}_2^{-1}$ e $\hat{K} = \hat{\Sigma}^{-1}$.

Ponendo $W = W_1 + W_2$ ed essendo $Tr(\hat{K}_i W_i) = n_i Tr(\hat{K}_i \hat{K}_i^{-1}) = n_i G$ e $Tr(\hat{K}W) = (n_1 + n_2)Tr(\hat{K}\hat{K}^{-1}) = (n_1 + n_2)G$ si ottiene

$$\Lambda = \prod_{i=1}^{2} \left(\frac{\det(\hat{K})}{\det(\hat{K}_i)} \right)^{\frac{n_i}{2}}$$

e quindi

$$-2\log(\Lambda) = \sum_{i=1}^{2} n_i \log \left(\frac{\det(\hat{K}_i)}{\det(\hat{K})} \right).$$

Sotto l'ipotesi nulla, se le matrici di varianza non sono shrinkate, si ha che la distribuzione asintotica di $T = -2\log(\Lambda)$ è un χ^2 con $r+G$ gradi di libertà, dove r è il numero di archi di $\mathcal{G}$. Dato il valore osservato del test T_0 l'ipotesi nulla si accetta se $Pr(\chi^2_{r+G} > T_0) > \alpha$. Nel caso invece in cui si procede alla stima attraverso metodi di shrinkage non è più possibile usare la distribuzione asintotica e si procede dunque con un approccio permutazionale. Esso sfrutta il fatto che se l'ipotesi nulla è vera ai fini del test gli esperimenti si possono scambiare tra una condizione e l'altra. Si ottengono quindi ω permutazioni casuali degli esperimenti ed altrettanti valori osservati del test $T_1, ..., T_\omega$. Il p-value è dato dunque da $\#(T_i \geq T_0)/\omega$ e l'ipotesi nulla è accettata se tale quantità e maggiore di una soglia α.

Un'altra ipotesi che è di interesse è la differenziale espressione tra i *pathway* nelle due diverse condizioni. Questo si traduce nella verifica d'ipotesi

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}.$$

Il test viene eseguito condizionatamente al fatto che l'ipotesi nulla nella (2.4) sia stata accettata. In tal caso il test può essere eseguito con una procedura esatta di analisi della varianza multivariata come ad esempio il T^2 di Hotelling. Nel caso l'ipotesi di omoschedasticità nella (2.4) sia rifiutata è possibile usare un usuale test per l'uguaglianza delle medie con diverse matrici di covarianza. Tale problema è detto problema di *Behrens-Fisher*, per approfondimenti si veda Anderson (2003).

Le stime di μ_1 e μ_2 per tali test sono date dalle medie campionarie, mentre le stime delle varianze sono calcolate col metodo di shrinking e con l'algoritmo IPS. Anche in questo caso è possibile adottare un approccio permutazionale.

2.3.3 IDENTIFICAZIONE DEI SIGNAL PATH RILEVENTI

Una volta che è stato identificato un *pathway* che si comporta in modo significativamente diverso tra due condizioni biologiche è di interesse scoprire quali porzioni di questo sono maggiormente associate a tale fenomeno.

Dato il grafo morale risultante da tale *pathway* la prima cosa da fare, se questo non è decomponibile, è di effettuare la triangolarizzazione. Dal grafo triangolarizzato si identificano le *clique* e si costruisce il relativo *junction tree*. Il concetto di *clique* non ha una facile interpretazione biologica ma è necessario per l'obiettivo che ci si è posto.

Per ogni *clique* viene testata quindi l'ipotesi di omoschedasticità come nel caso descritto per il test globale per l'intero *pathway*, con il vantaggio che non vi è bisogno di imporre i valori nulli nelle stime delle matrici di covarianza in corrispondenza degli archi mancanti con l'algoritmo IPS in quanto in una *clique* ogni nodo è collegato con tutti gli altri nodi della stessa. I p-value risultanti da tale operazione saranno i pesi w dei nodi del *junction tree*. Un peso viene considerato significativo se è minore di un determinato α.

Si passa poi all'identificazione di tutti i cammini massimali del *junction tree*, ovvero quei cammini che iniziano col nodo radice e terminano con una foglia. Per ogni singolo cammino massimale identificato si selezionano le sue porzioni distinte di maggior lunghezza composte da *clique* significative

consecutive o separate al massimo da una sola *clique* non significativa. Tali porzioni vengono dette *sub-paths*.

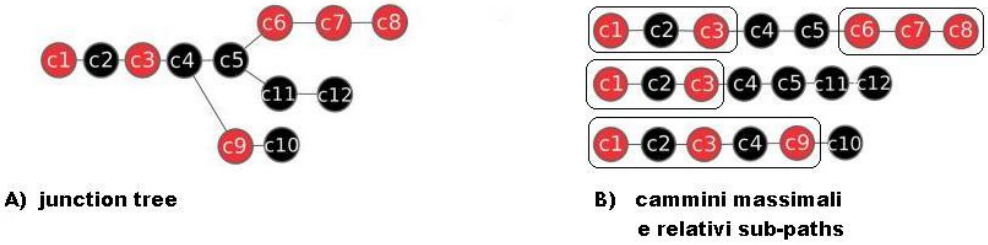

Figura 2.3: In A) è mostrato un esempio di *junction tree*, mentre in B) si mostrano i suoi cammini massimali. Le clique marcate in rosso rappresentano le *clique* significative, mentre i raggruppamenti in B) rappresentano i *sub-paths*.

Si passa dunque al calcolo della rilevanza di ogni *sub-path*. Sia L_j la lunghezza del *sub-path* j, con $j = 1, ..., J$ e dati i pesi w_{ij} per ogni *clique* i nel *sub-path* j, con $i = 1, ..., L_j$, nel rispetto dell'ordine delle *clique* nel *sub-path*, per ogni *clique* i nel *sub-path* j si calcola la seguente quantità

$$S_{ij} = \sum_{k=1}^{i} \delta_{kj}, \quad i = 1, ..., L_j$$

dove si definisce δ_{kj} come

$$\delta_{kj} = \begin{cases} -\log(w_{kj}) & w_{kj} < \alpha \\ log(1 - w_{kj}) & w_{kj} \geq \alpha \end{cases}.$$

Si definisce dunque la rilevanza R_j relativa del *sub-path* j come il massimo valore tra $S_{1j}, ..., S_{L_j j}$. Per rendere comparabili le rilevanze di *sub-paths* di lunghezza diversa si definisce per ogni *sub-path* la rilevanza standardizzata seguente

$$SR_j = \frac{R_j \cdot m_j}{L_j}$$

dove m_j è la posizione dell'elemento R_j all'interno della lista $S_{1j}, ..., S_{L_j j}$. Si può quindi ottenere per ogni cammino massimale il *sub-path* con la massima

rilevanza standardizzata che viene detto *signal-path* rilevante.

Nella maggior parte dei casi comunque i cammini massimali e quindi i *sub-paths* e i *signal-paths* che si ricavano sono molto sovrapposti e quindi hanno molti geni in comune. Si definisce dunque la seguente misura di dissimilarità tra due generici *sub-path* A e B

$$d(A,B) = \begin{cases} \frac{|A-B|}{|A|} & |A-B| \leq |B-A| \\ \frac{|B-A|}{|B|} & |A-B| > |B-A| \end{cases}$$

dove $|A|$ è l'insieme dei geni corrispondenti al *sub-path* A, $|B|$ è l'insieme dei geni corrispondenti al *sub-path* B ed $|A-B|$ e $|B-A|$ sono le cardinalità di una differenza di insiemi. I *sub-paths* che hanno una misura di dissimilarità inferiore ad un dato ϵ vengono quindi collassati, dove con collasso si intende tenere il *sub-path* con la rilevanza relativa più alta.

2.4 SCOPO DELLA TESI

Lo scopo della tesi è quello di utilizzare i metodi esposti in questo capitolo per identificare non solo dei geni differenzialmente espressi, ma anche dei microRNA differenzialmente espressi e dei geni differenzialmente metilati. Il fine è quello di cercare delle corrispondenze a posteriori tra i geni codificanti proteine differenzialmente espressi ed i livelli di metilazione genomica e di espressione dei microRNA, con l'obiettivo di individuare dei possibili meccanismi biologici che possano essere importanti per una migliore comprensione dell'evoluzione del tumore all'ovaio a livello di biologia molecolare.

Capitolo 3

PRESENTAZIONE DEI DATI

3.1 LA PATOLOGIA

Il carcinoma ovarico è un tumore che colpisce le ovaie, due organi delle dimensioni di circa tre centimetri situati uno a destra e uno a sinistra dell'utero al quale sono connessi tramite le tube di Falloppio. Le ovaie sono deputate alla produzione di ormoni sessuali femminili e di ovociti, ovvero le cellule riproduttive femminili: ogni mese, quando la donna è fertile e non è in stato di gravidanza, le ovaie producono un ovocita che si muove verso l'utero per essere fecondato.

Nel mondo occidentale, tra i tumori ginecologici, il carcinoma ovarico è il secondo per frequenza ed il primo come causa di morte. In Italia, secondo le stime del 2012 del Registro Tumori, il tumore dell'ovaio colpisce in media 4.490 donne ogni anno. Considerando le altre forme tumorali esso è al nono posto per frequenza, costituendo il 2,9% di tutte le diagnosi di tumore. In Europa rappresenta il 5% di tutti i tumori femminili. Risulta più frequente nella popolazione caucasica, nei Paesi dell'Europa nord occidentale e negli USA, assai meno frequente nei Paesi asiatici, africani, sudamericani[1].

Nel 90% dei casi il carcinoma ovarico ha origine dalle cellule epiteliali, ovvero le cellule che ricoprono superficialmente le ovaie. Si dice quindi che il tumore è epiteliale. Nei restanti casi il tumore può svilupparsi dalle cellule

[1]Dati forniti dall'AIRC, l'associazione italiana per la ricerca sul cancro.

germinali, che sono le cellule che producono gli ovociti, o dalle cellule del tessuto dello stroma gonadico, che è il tessuto di sostegno dell'ovaio. In tali casi il tumore viene detto rispettivamente germinale e stromale.

Il cancro all'ovaio è dovuto alla proliferazione incontrollata delle cellule cancerose nell'organismo. La classificazione più comunemente utilizzata per descrivere lo stato di diffusione del tumore è quella della FIGO (Federazione Internazionale di Ginecologia Ostetricia) ed è descritta nella Tabella 3.1. La Figura 3.1 invece è stata inserita per dare una visione schematica più immediata.

Il tumore all'ovaio nei primi due stadi non dà sintomi facilmente riconoscibili ma solo vaghi dolori addominali o pelvici e senso di gonfiore, per cui la diagnosi viene molto spesso effettuata in fase avanzata, ovvero quella corrispondente agli stadi III e IV.

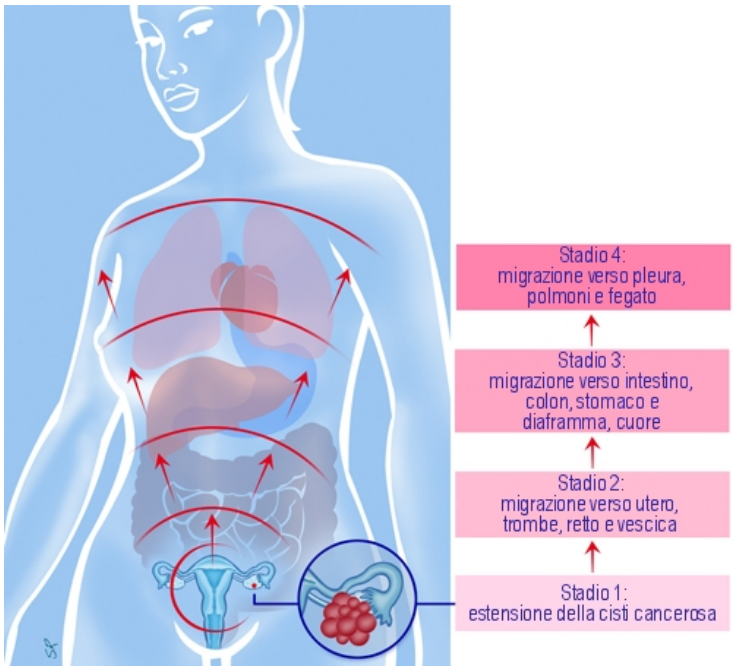

Figura 3.1: Rappresentazione schematica semplificata per visualizzare l'evoluzione e l'estensione del carcinoma ovarico agli altri organi.

Stadio I	Il carcinoma è limitato a un ovaio o a entrambi. Esso si sviluppa all'interno di una ciste ospite per poi romperne la parete ed estendersi all'esterno nell'ovaio (vegetazioni).
Stadio II	Le cellule cancerose si estendono all'interno della cavità addominale circondata dal peritoneo, pur permanendo all'interno della pelvi (porzione inferiore del peritoneo). Possono così intaccare l'utero, le trombe, il sacco rettale e la vescica.
Stadio III	Le cellule cancerose si estendono verso l'alto, all'interno dell'addome, in direzione dell'intestino, del colon, dello stomaco e del diaframma. Una volta che il peritoneo viene attaccato dalle cellule cancerose, esso produce liquido (ascite) che si deposita nell'addome. Le cellule cancerose possono raggiungere e colonizzare i linfonodi localizzati a livello dei vasi cardiaci, dell'aorta e della vena cava.
Stadio IV	Il cancro si diffonde al di là dell'addome e raggiunge la pleura (tessuto che circonda i polmoni) dove produce un liquido detto pleurite e si sposta quindi verso altri organi quali i polmoni o il fegato. Si parla allora di metastasi, ovvero estensione delle cellule cancerose ad altri organi, a distanza.

Tabella 3.1: Stadiazione del tumore all'ovaio secondo la classificazione FIGO.

3.2 TGCA

The Cancer Genome Atlas (TCGA) è un progetto nato nel 2005 per raccogliere informazioni genetiche sui soggetti malati di cancro. Esso viene fondato da due importanti istituti statunitensi: il National Cancer Institute (NCI) e il National Human Genome Research Institute (NHGRI). Sono gli scienziati e i manager di questi istituti a gestire questo progetto per la lotta contro il cancro.

Nel 2006 viene lanciato un progetto pilota triennale il cui scopo principale è di caratterizzare tre tipi di cancro umano: il glioblastoma multiforme (un tumore del cervello molto aggressivo), il cancro ai polmoni e il cancro all'ovaio.

L'obiettivo di tale progetto pilota era quello di dimostrare che le avanzate tecnologie biologiche a disposizione potevano essere utilizzate da un team di scienziati provenienti da vari istituti per generare conclusioni biologiche e statistiche significative provenienti dai data set raccolti.

Nel 2009 il progetto è stato esteso alla cosidetta 'phase II', che si prefigge di completare la caratterizzazione genomica e l'analisi di sequenziamento di sequenziamento di altri 20 tipi di tumori. Tale fase al momento della stesura di questa tesi è ancora in atto.

Il progetto organizza e raccoglie solitamente campioni provenienti da 500 a 600 pazienti (molti di più di quelli degli usuali studi genetici) e li analizza con differenti tecniche, tra le quali tecniche per lo studio dei profili di espressione genica, per il profilo dei microRNA e per il profilo della metilazione del DNA.

I dati oggetto di questa tesi provengono da tale progetto e sono descritti al paragrafo seguente.

3.3 I DATI

I dati forniti da TGCA relativi al cancro ovarico che si andranno a prendere in considerazione per le analisi sono: i) un gruppo di 594 esperimenti per rilevare il livello di espressione dei geni codificanti effettuati con microarray a singolo canale della Affymetrix, ii) un gruppo di 587 esperimenti per rilevare il livello di espressione dei microRNA effettuati con microarray a singolo canale della Agilent, iii) un gruppo di 605 esperimenti per la rilevazione della intensità di metilazione ϕ^m e non di metilazione ϕ^{nm} dei geni codificanti.

	I	II	III	IV	Stadio non registrato	TOTALE
Geni codificanti	16	30	450	84	14	594
MicroRNA	16	29	455	85	2	587
Metilazione	17	30	454	86	18	605

Tabella 3.2: Tabella di frequenza degli esperimenti (repliche comprese) condizionata al tipo di esperimento effettuato e allo stadio FIGO su cui questo è eseguito.

In ogni gruppo di esperimenti su una singola paziente affetta da cancro ovarico viene effettuato un solo esperimento, tranne che per 16, 8 e 21 pazienti relative rispettivamente alle serie di esperimenti per i geni codificanti, per i microRNA e per la metilazione su cui sono state effettuate due repliche tecniche degli stessi esperimenti.

La Tabella 3.2 mostra per ogni gruppo di esperimenti quanti di questi sono stati effettuati su un determinato stadio FIGO. Nel seguito si paragoneranno le due condizioni biologiche 'stadio iniziale' e 'stadio avanzato' accorpando rispettivamente lo stadio I con lo stadio II e lo stadio III con lo stadio IV. Si nota che tale raggruppamento risulta sbilanciato, essendo gli esperimenti relativi allo stadio avanzato molto più numerosi di quelli dello stadio iniziale, tuttavia esso è l'unico che ha senso considerare dal punto di vista biologico in quanto il passaggio dallo stadio II allo stadio III è un punto chiave che rappresenta il passaggio da una condizione di tumore ancora localizzata ad una condizione di tumore estesa molto invasiva.

Va inoltre specificato che i geni codificanti degli esperimenti di espressione e metilazione sono identificati con un codice numerico detto entrez mentre i microRNA sono identificati con un loro specifico codice alfanumerico.

Capitolo 4

ANALISI DEI DATI

4.1 PULITURA E NORMALIZZAZIONE

Il primo passo consiste nell'eliminare tutti gli esperimenti relativi alle pazienti di cui non è nota la condizione clinica. Si eliminano inoltre tutti gli esperimenti relativi alle pazienti che non hanno effettuato tutte e tre le serie di esperimenti (geni codificanti, microRNA, metilazione).

Vengono poi calcolati i *β-values* degli esperimenti di metilazione. Si ottengono così tre matrici (una per ogni serie di esperimenti) dove ogni riga rappresenta una rilevazione e ogni colonna rappresenta un esperimento o una replica tecnica. In ogni matrice ottenuta si collassano con la media i valori delle colonne che rappresentano repliche tecniche di uno stesso esperimento. Sulle tre matrici poi si fa un lavoro di pulizia per eliminare rilevazioni mal annotate o poco affidabili. Vengono inoltre tolte dalla matrice dei geni codificanti le rilevazioni relative ai probe di controllo di Affymetrix, mentre sulla matrice dei microRNA, siccome Agilent replica ogni spot 16 volte per avere una maggiore qualità, si collassano le rilevazioni relative a spot dello stesso tipo col valore mediano.

Fatto ciò si effettua la normalizzazione quantile sulla matrice dei geni codificanti e sulla matrice dei microRNA, i livelli di metilazione invece sono già normalizzati. La Figura 4.1 mostra un esempio dell'effetto della normalizzazione quantile. Si calcolano poi i logaritmi dei valori di queste due matrici e

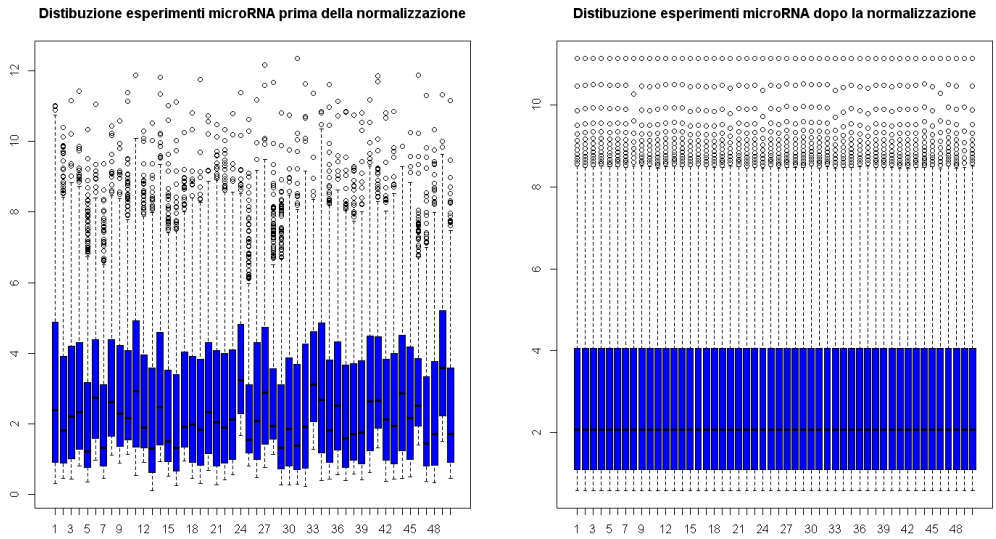

Figura 4.1: La figura mostra l'effetto che ha la normalizzazione quantile sulle distribuzioni degli esperimenti. Si sono presi come esempio solo 50 esperimenti microRNA per motivi grafici. Inoltre, data la presenza di valori esterni ai baffi molto elevati, in entrambi i grafici valori dei boxplot sono su scala logaritmica.

il logit dei β-*values* della matrice di metilazione.

Si ottengono dunque una matrice di dati relativa ai geni codificanti di dimensione 12067×554, una matrice dei dati relativa ai microRNA di dimensione 799×554 ed una matrice di dati relativa ai livelli di metilazione di dimensione 13415×554.

4.2 DIFFERENZIALE ESPRESSIONE E METILAZIONE

Ora che sono state ottenute le matrici dei dati, si va ad applicare il test bayesiano empirico per lo studio della differenziale espressione e metilazione.

Il calcolo delle statistiche test t moderate per la matrice dei geni codificanti porta ad avere un p-value per ogni gene. Prima di passare al calcolo dei q-value per i dati di espressione genica si costruisce l'istogramma di frequenza dei p-value (Figura 4.2) per visualizzare graficamente la soglia λ oltre la quale

la distribuzione dei p-value è considerata uniforme. Viene quindi fissato un λ pari a 0.55 che porta a stimare π_0 con $\hat{\pi}_0 = 3967/12067 \doteq 0.33$. Si ottiene dunque una lista di q-value che con una soglia $\alpha = 0.05$ porge la seguente lista di 79 geni differenzialmente espressi:

10005	10663	10671	10753	116984	1263	1287	131	135948
1815	1846	201229	2139	22797	22890	22953	231	23466
240	25806	25822	25987	26003	26521	27136	2731	27440
2780	2916	2920	3060	3069	317762	3291	3674	369
420	4660	4674	4880	4891	50837	51157	51299	51555
5199	5394	5440	54733	55256	55259	5546	5573	55833
55911	5741	5754	60385	6367	63931	64693	6605	6777
715	7181	7371	7866	79874	79977	80153	8170	83443
8742	9379	9414	9439	9529	9813	9848		

I q-value dei geni codificanti differenzialmente espressi sono mostrati in appendice.

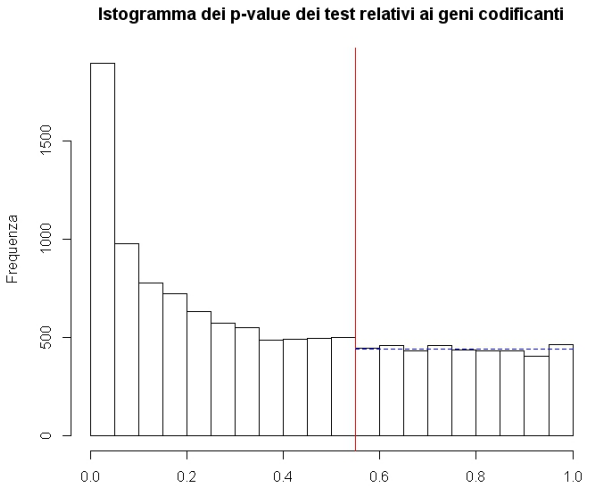

Figura 4.2: Istogramma di frequenza dei p-value relativi ai geni codificanti. La linea rossa rappresenta la soglia $\lambda = 0.55$ oltre la quale la distribuzione dei p-value sembra uniforme.

Si passa poi al calcolo delle statistiche test t moderate per la matrice dei microRNA e dei relativi p-value. La stima di $\lambda^{(m)}$ (Figura 4.3) è pari a 0.8 e porta a stimare $\pi_0^{(m)}$ con $\hat{\pi}_0^{(m)} = 115/799 \doteq 0.14$. Si ottiene quindi una lista di q-value che con una soglia $\alpha = 0.05$ porge la seguente lista di microRNA differenzialmente espressi con i relativi q-value:

microRNA	q-value
ebv-miR-BART14*	0.042834136
hsa-let-7a-3p	0.005570097
hsa-let-7i-3p	0.042834136
hsa-miR-199a-5p	0.043223706
hsa-miR-199b-5p	0.043223706
hsa-miR-200c-3p	0.043223706
hsa-miR-204-5p	0.020957602
hsa-miR-22-5p	0.022916912
hsa-miR-506-3p	0.042834136
hsa-miR-509-3-5p	0.043223706

Si nota che nella lista dei microRNA differenzialmente espressi individuati vi è la presenza di un microRNA di origine virale: ebv-miR-BART14* appartenente al virus di Epstein-Barr. Allo stato attuale non è ben noto come i microRNA virali agiscano e come interpretarli dal punto di vista biologico, si decide dunque di rimuoverlo dalle analisi successive.

Si calcolano infine le statistiche test t moderate per la matrice dei livelli di metilazione e dei relativi p-value. La stima di $\lambda^{(M)}$ (Figura 4.4) è pari a 0.55 e porta a stimare $\pi_0^{(M)}$ con $\hat{\pi}_0^{(M)} = 4243/13415 \doteq 0.31$. Si ottiene quindi una lista di q-value ma non si trovano geni differenzialmente metilati per i livelli di significatività $\alpha = 0.05$ e $\alpha = 0.1$. Questo non significa che nella realtà essi non ci siano ma solo che il metodo non è sufficientemente potente da rilevarli.

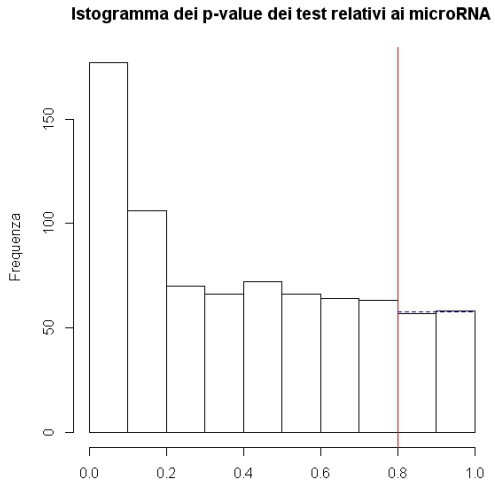

Figura 4.3: Istogramma di frequenza dei p-value relativi ai microRNA. La linea rossa rappresenta la soglia $\lambda = 0.8$ oltre la quale la distribuzione dei p-value sembra uniforme.

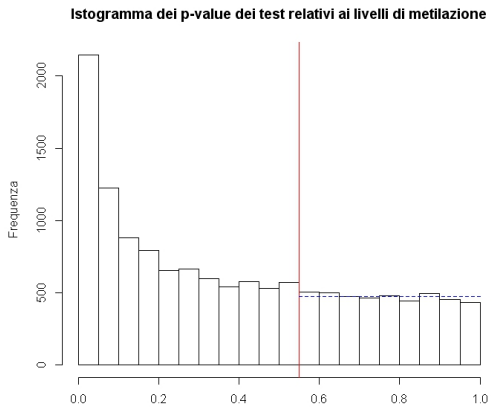

Figura 4.4: Istogramma di frequenza dei p-value relativi ai livelli di metilazione. La linea rossa rappresenta la soglia $\lambda = 0.55$ oltre la quale la distribuzione dei p-value sembra uniforme.

4.3 ANALISI DEI PATHWAY

In questo paragrafo si andranno ad applicare dei modelli grafici gaussiani alle matrici dei livelli di espressione e dei livelli di metilazione dei geni codificanti per ognuno dei 236 *pathway* della libreria KEGG.

Per fare questo si convertono i *pathway* in grafi di soli geni. Per ogni grafo di geni relativo ad un dato *pathway* si devono poi estrarre due sottografi, uno per ogni matrice di dati. I nodi dei sottografi sono dati dall'intersezione tra i geni del grafo di soli geni originario e dei geni presenti nelle rispettive matrici. Ogni sottografo viene poi trasformato in un DAG e moralizzato per costruire i rispettivi modelli.

Per ogni modello si effettua un test con 100 permutazioni per l'ipotesi di omoschedasticità e condizionatamente a questo con un livello $\alpha = 0.05$ un test per l'uguaglianza delle medie utilizzato sempre 100 permutazioni.

Ottenuti i p-value di ogni test è di interesse cercare quei *pathway* che risultano sregolati, dal punto di vista dell'espressione e della metilazione genica, sia in media che in varianza. Con una soglia $\alpha = 0.01$ applicata su ogni test i *pathway* che risultano sregolati sono i seguenti:

- Measles, il *pathway* relativo ad una malattia che attacca il sistama respiratorio, il sistama immunitario e la pelle;
- Prostate cancer, il *pathway* relativo al cancro alla prostata;
- Acute myeloid leukemia, il *pathway* relativo alla leucemia.

Identificati questi *pathway* è ora di interesse capire quali geni in essi sono maggiormente associati al passaggio da stadio iniziale a stadio avanzato e vedere come questo si leghi con il livello di metilazione e con l'azione dei microRNA. Per fare questo vengono calcolati i *signal-paths* relativi a tali *pathway* sia per il livello di espressione dei geni codificanti che per il livello di metilazione. Per il calcolo della significatività delle *clique* si utilizza un approccio permutazionale (100 permutazioni) considerandole significative con una soglia $alpha = 0.05$. Per i microRNA invece si farà riferimento

ai geni target della lista identificata col metodo bayesiano empirico (si veda l'appendice per la lista completa di tutti i geni target della lista).

Iniziando ad esplorare il grafo di soli geni del *pathway* measles si nota che esso risulta composto da molte componenti sconnesse, questo forse è dovuto a una cattiva traduzione del *pathway* in grafo di soli geni. Si ha inoltre che i *signal-path* più rilevanti in espressione genica e metilazione non hanno geni in comune e non ci sono nemmeno target dei microRNA nel *signal-path* dei geni codificanti. Si decide dunque di non mostrare la rappresentazione del grafo di geni di tale *pathway* perché ritenuta poco interessante.

Si passa poi a esplorare il grafo di geni relativo al cancro alla prostata.

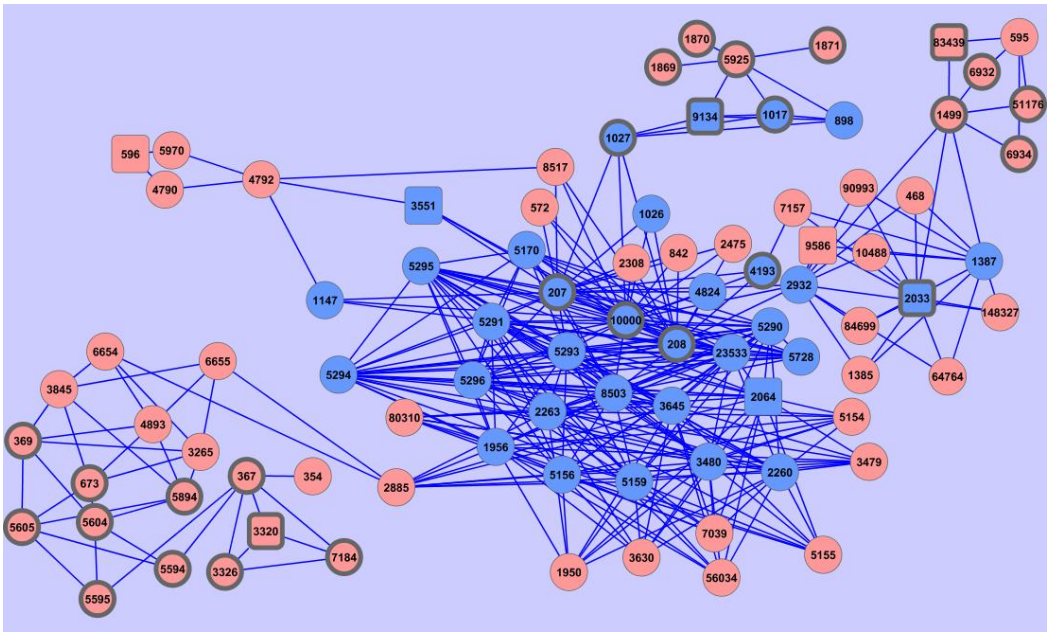

Figura 4.5: La figura mostra il grafo di soli geni relativo al *pathway* prostate cancer. I nodi in blu rappresentano i geni appartenenti al *signal-path* più rilevante relativo all'espressione genica mentre i nodi cerchiati in grassetto rappresentano i geni appartenenti al *signal-path* più rilevante relativo al livello di metilazione. I nodi con cornice quadrata invece rappresentano i geni target di uno o più microRNA differenzialmente espressi.

La Figura 4.5 mostra che un'ampia porzione dei geni del *pathway* relativo al cancro alla prostata risulta sregolato in espressione dal passaggio alla fase avanzata del tumore all'ovaio. Si nota inoltre che il *signal-path* più ri-

levante in metilazione è in parte sovrapposto a quello dell'espressione genica, rilevando quindi una relazione tra espressione genica e metilazione che con il metodo bayesiano empirico non era stata colta. Si ha anche che i seguenti geni risultano sregolati a causa dei microRNA:

- 3551 sregolato a causa di hsa-miR-199a-5p e di hsa-miR-200c-3p;
- 2064 sregolato a causa di hsa-miR-199b-5p e di hsa-miR-199a-5p;
- 2033 sregolato a causa di hsa-miR-204-5p e di hsa-miR-200c-3p;
- 9134 sregolato a causa di hsa-miR-200c-3p.

È interessante notare inoltre come i geni 2033 e 9134 risultino sregolati a causa dell'effetto congiunto della metilazione e dei microRNA.

Si esplora ora il grafo di geni relativo alla leucemia. La Figura 4.6 che il *signal-path* più rilevante relativo all'espressione genica mappato sul grafo è abbastanza esteso e che il *signal-path* più rilevante relativo al livello di metilazione mappato, pur non essendo molto esteso, risulta quasi totalmente sovrapposto a quest'ultimo. Si nota inoltre che esso contiene nuovamente il gene 3551 target dei microRNA hsa-miR-199a-5p e hsa-miR-200c-3p.

Si considerano ora i *pathway* sregolati solo in media sia per l'espressione che per la metilazione genica con una significatività pari ad $\alpha = 0.01$ (si veda l'appendice per la lista e i p-value). Tra di essi se ne notano due di molto famosi nello studio dei tumori che si decide di esaminare: apoptosis e p53 signaling pathway.

Apoptosis è il *pathway* relativo alla apoptosi, una forma di morte cellulare programmata che dagli inizi degli anni novanta ad oggi ha visto un forte incremento del suo studio. Questo è dovuto al fatto che oltre alla sua importanza come fenomeno biologico ha acquisito un enorme valore medico in quanto processi difettosi di apoptosi riguardano numerose malattie. Una eccessiva attività apoptotica può causare disordini da perdita di cellule (si vedano ad esempio alcune malattie neurodegenerative, come la malattia di Parkinson), mentre un'apoptosi carente può implicare una crescita cellulare incontrollata, meccanismo alla base della formazione di masse tumorali.

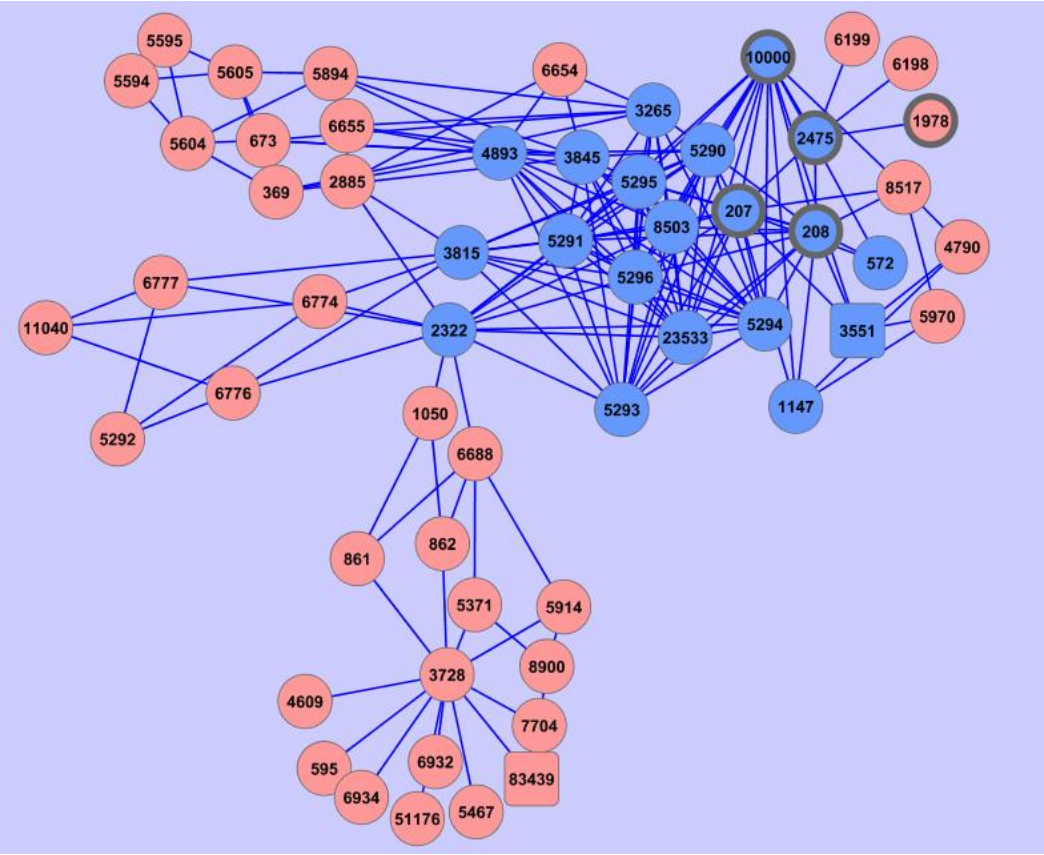

Figura 4.6: La figura mostra il grafo di soli geni relativo al *pathway* acute myeloid leukemia. Per il significato della simbologia utilizzata si faccia riferimento alla didascalia della Figura 4.5.

Il grafo di soli geni relativo al *pathway* dell'apoptosi è rappresentato nella Figura 4.7. Si nota che i geni che in esso risultano differenzialmente espressi sono quasi totalmente anche differenzialmente metilati. Si nota anche che quattro di essi sono target di microRNA:

- 3551 sregolato a causa di hsa-miR-199a-5p e di hsa-miR-200c-3p;
- 329 sregolato a causa di hsa-miR-204-5p;
- 331 sregolato a causa di hsa-miR-200c-3p;
- 596 sregolato a causa di hsa-miR-200c-3p.

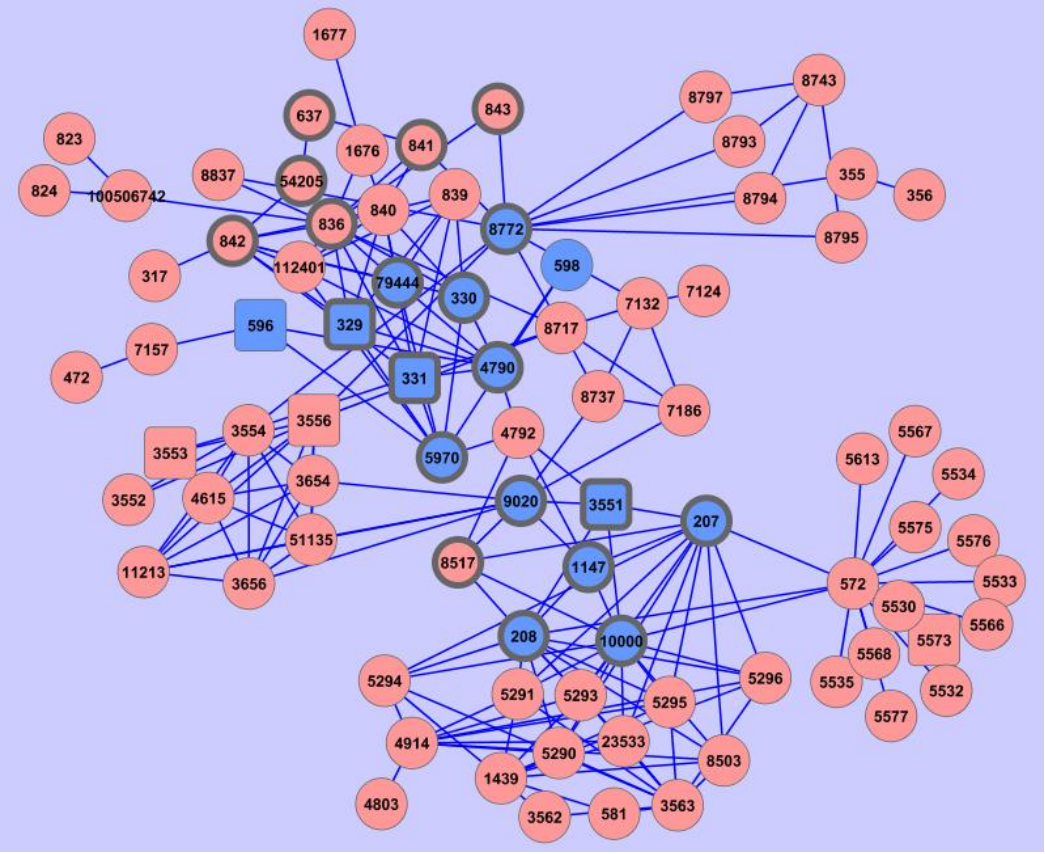

Figura 4.7: La figura mostra il grafo di soli geni relativo al *pathway* apoptosis. Per il significato della simbologia utilizzata si faccia riferimento alla didascalia della Figura 4.5.

Si ritrova dunque il gene 3551 che si era identificato precedentemente nei *pathway* della prostata e della leucemia. Si ritiene dunque il fatto che tale gene sia target di due microRNA rilevante. Dal punto di vista biologico si ha che il gene 3551 codifica per una proteina il cui compito è quello di effettuare una fosforilazione (ovvero va ad addizionare un gruppo fosfato PO_4^{3-}) all'inibitore del complesso inibitore/NF-κB, causandone quindi la distruzione ed attivando il complesso proteico NF-κB. Il complesso una volta attivato protegge le cellule che in condizioni normali sarebbero uccise dal meccanismo dell'apoptosi, consentendone così la prolificazione.

Si decide ora di prendere in considerazione l'intersezione dei geni sregolati in espressione e metilazione dei *signal-paths* più rilevanti dei *pathway* relativi alla prostata, alla leucemia e all'apoptosi. Tale intersezione è data dai geni 207, 208 e 10000. Questo gruppo di geni ha un forte significato biologico in quanto essi codificano rispettivamente per le proteine AKT1, AKT2 e AKT3. Queste tre proteine costituiscono la famiglia proteica AKT, conosciuta anche col nome di Protein Kinase B (PKB), che gioca un ruolo fondamentale in molti processi cellulari: il metabolismo del glucosio, la trascrizione, l'apoptosi, la prolificazione cellulare e la migrazione delle cellule. Si ritiene perciò sensato supporre che la metilazione di tali geni giochi un ruolo fondamentale nel passaggio dalla fase iniziale più localizzata alla fase avanzata, e quindi più diffusa, del tumore all'ovaio.

Si prende ora in considerazione p53 signaling pathway. Esso è il *pathway* relativo alla proteina p53, anche conosciuta come proteina tumorale 53 (trascritta dal gene 7157 o TP53 con notazione symbol). Questa è un fattore di trascrizione che regola il ciclo cellulare e ricopre la funzione di soppressore tumorale. La sua funzione è particolarmente importante negli organismi pluricellulari per sopprimere i tumori nascenti. La p53 è stata descritta come 'il guardiano del genoma' riferendosi al suo ruolo di preservazione della stabilità attraverso la prevenzione delle mutazioni. Deve il suo nome alla semplice massa molecolare: pesa infatti 53 kDa. La proteina p53 è stata identificata nel 1979 ma il suo carattere di gene soppressore tumorale è stato rivelato solamente nel 1989 da Bert Vogelstein della Johns Hopkins School of Medicine.

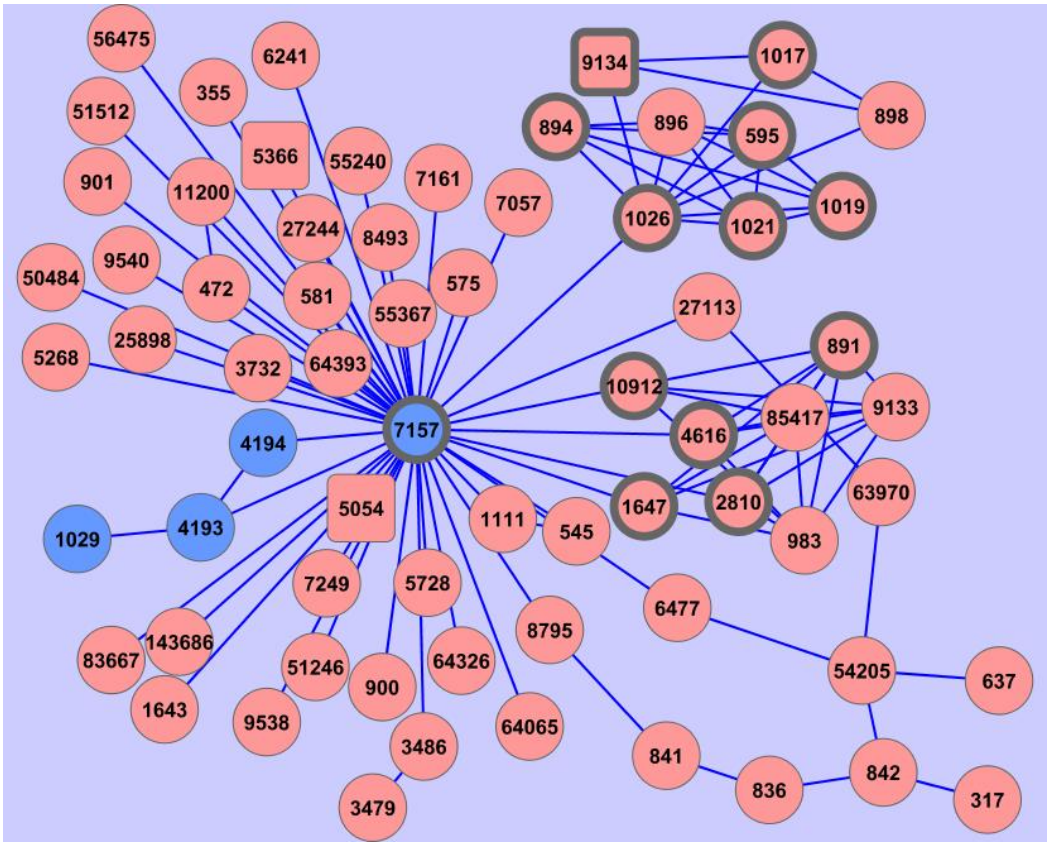

Figura 4.8: La figura mostra il grafo di soli geni relativo al *pathway* p53 signaling pathway. Per il significato della simbologia utilizzata si faccia riferimento alla didascalia della Figura 4.5.

La sua importanza è tale che nel 1993 la prestigiosa rivista Sience l'ha eletta molecola dell'anno.

Il grafo di soli geni relativo a tale *pathway* è mostrato nella Figura 4.8. In esso si nota come il gene 7157 abbia una posizione centrale e che da esso partano numerosi archi. I geni dei due *signal-paths* più rilevanti in espressione e metilazione genica coincidono proprio sul gene che codifica la proteina p53. Si ritiene dunque sensato supporre che metilazione del gene 7157 porti a una bassa espressione di questo oncosoppressore favorendo quindi il passaggio da stadio iniziale localizzato a stadio avanzato diffuso del tumore all'ovaio.

Capitolo 5

CONCLUSIONI

Dopo una fase iniziale di creazione e normalizzazione delle matrici di dati si è passati allo studio della differenziale espressione e metilazione utilizzando il metodo bayesiano empirico che però non ha identificato geni che siano contemporaneamente differenzialmente espressi e differenzialmente metilati data la mancata rilevazione di questi ultimi. Il metodo però è stato utile per identificare una lista di microRNA differenzialmente espressi da integrare nelle analisi successive.

Si è poi effettuata una analisi basata sui modelli grafici gaussiani utilizzando i *pathway* della libreria KEGG. Questo ha portato all'individuazione di tre *pathway* sregolati sia in media che in varianza sia per l'espressione che per la metilazione genica con una soglia $\alpha = 0.01$. Si sono poi andati a considerare i *signal-path* più rilevanti in espressione e metilazione. L'assenza di nodi in comune tra i due *signal-path* in uno di questi *pathway* l'ha portato ad essere scartato. Ci si è quindi concentrati sugli altri due che erano relativi a due diverse forme tumorali: il cancro alla prostata e la leucemia.

Si è poi deciso di considerare i *pathway* sregolati in media sia per l'espressione che per la metilazione genica con una soglia $\alpha = 0.01$. Tra di essi si è scelto di analizzarne due di molto famosi in ambito tumorale, quello relativo alla apoptosi e quello relativo alla proteina p53.

Lo studio del *pathway* dell'apoptosi ha portato ad identificare la sregolazione di due fenomeni biologici tra la condizione iniziale circoscritta e la

condizione avanzata estesa di cancro ovarico comuni anche ai *pathway* delle forme tumorali che erano state prese in considerazione precedentemente. Sembra dunque fondato affermare che la differenziale espressione dei microRNA hsa-miR-199a-5p e hsa-miR-200c-3p che hanno come target il gene 3551 e che la metilazione del gruppo di geni 207, 208 e 10000 che codificano per le proteine della famiglia proteica AKT influenzino il passaggio da stadio iniziale a stadio avanzato del tumore ovarico.

Lo studio del *pathway* relativo alla proteina p53 invece suggerisce che tale passaggio sia anche influenzato dalla metilazione del gene 7157 che codifica tale proteina.

Appendice A

TABELLE

Questa appendice contiene tre tabelle relative alle analisi effettuate:

- la Tabella A.1 fornisce i q - value dei geni codificanti differenzialmente espressi con $\alpha = 0.05$ identificati col test bayesiano empirico;
- la Tabella A.2 fornisce i geni target dei microRNA differenzialmente espressi con $\alpha = 0.05$ identificati col test bayesiano empirico;
- la Tabella A.3 fornisce i p-value dei *pathway* significativi in media per l'espressione e la metilazione dei geni codificanti con $\alpha = 0.05$ relativi ai modelli grafici gaussiani.

ENTREZ	q-value	ENTREZ	q-value	ENTREZ	q-value
10005	0.04061171	2780	0.04061171	55911	0.03441678
10663	0.03678453	2916	0.04444183	5741	0.04061171
10671	0.04061171	2920	0.03283615	5754	0.04061171
10753	0.03083825	3060	0.02668114	60385	0.03083825
116984	0.03283615	3069	0.04061171	6367	0.03678453
1263	0.04674866	317762	0.04061171	63931	0.03083825
1287	0.04590066	3291	0.04674866	64693	0.04674866
131	0.01117478	3674	0.03597199	6605	0.03409274
135948	0.03083825	369	0.04061171	6777	0.03597199
1815	0.04061171	420	0.04061171	715	0.03678453
1846	0.04061171	4660	0.04593056	7181	0.04061171
201229	0.01117478	4674	0.04061171	7371	0.03283615
2139	0.03283615	4880	0.04061171	7866	0.04674866
22797	0.02668114	4891	0.03083825	79874	0.03678453
22890	0.03083825	50837	0.04061171	79977	0.04061171
22953	0.03441678	51157	0.03678453	80153	0.04061171
231	0.04674866	51299	0.03083825	8170	0.03678453
23466	0.03083825	51555	0.04579124	83443	0.04725282
240	0.04061171	5199	0.03083825	8742	0.04061171
25806	0.04674866	5394	0.03678453	9379	0.04593056
25822	0.04061171	5440	0.04061171	9414	0.04674866
25987	0.04674866	54733	0.03083825	9439	0.04319947
26003	0.04674866	55256	0.01205879	9529	0.03083825
26521	0.04674866	55259	0.04061171	9813	0.04444183
27136	0.04061171	5546	0.04061171	9848	0.04061171
2731	0.04061171	5573	0.03461392		
27440	0.04061171	55833	0.03083825		

Tabella A.1: q - value dei geni codificanti differenzialmente espressi.

MiRNA	ENTEZ	MiRNA	ENTEZ	MiRNA	ENTEZ
hsa-miR-199a-5p	6662	hsa-miR-200c-3p	5366	hsa-miR-204-5p	3037
hsa-miR-199a-5p	6595	hsa-miR-200c-3p	4908	hsa-miR-204-5p	3589
hsa-miR-199a-5p	2146	hsa-miR-200c-3p	1902	hsa-miR-204-5p	3553
hsa-miR-199a-5p	3551	hsa-miR-200c-3p	1909	hsa-miR-204-5p	3556
hsa-miR-199a-5p	57018	hsa-miR-200c-3p	387	hsa-miR-204-5p	3576
hsa-miR-199a-5p	3976	hsa-miR-200c-3p	27252	hsa-miR-204-5p	3690
hsa-miR-199a-5p	3726	hsa-miR-200c-3p	5789	hsa-miR-204-5p	3720
hsa-miR-199a-5p	10001	hsa-miR-200c-3p	63916	hsa-miR-204-5p	4074
hsa-miR-199a-5p	4204	hsa-miR-200c-3p	55914	hsa-miR-204-5p	10982
hsa-miR-199a-5p	2114	hsa-miR-200c-3p	22884	hsa-miR-204-5p	4653
hsa-miR-199a-5p	780	hsa-miR-200c-3p	55697	hsa-miR-204-5p	5327
hsa-miR-199a-5p	1906	hsa-miR-200c-3p	83439	hsa-miR-204-5p	5329
hsa-miR-199a-5p	4296	hsa-miR-200c-3p	9770	hsa-miR-204-5p	57403
hsa-miR-199a-5p	4089	hsa-miR-200c-3p	3215	hsa-miR-204-5p	10966
hsa-miR-199a-5p	6783	hsa-miR-200c-3p	54453	hsa-miR-204-5p	10955
hsa-miR-199a-5p	27201	hsa-miR-200c-3p	8462	hsa-miR-204-5p	27230
hsa-miR-199a-5p	2064	hsa-miR-200c-3p	989	hsa-miR-204-5p	5054
hsa-miR-199a-5p	7374	hsa-miR-200c-3p	6464	hsa-miR-204-5p	6938
hsa-miR-199a-5p	857	hsa-miR-200c-3p	55659	hsa-miR-204-5p	6925
hsa-miR-199a-5p	23411	hsa-miR-200c-3p	3609	hsa-miR-204-5p	79071
hsa-miR-199b-5p	3280	hsa-miR-200c-3p	256471	hsa-miR-204-5p	860
hsa-miR-199b-5p	6418	hsa-miR-200c-3p	4508	hsa-miR-204-5p	6659
hsa-miR-199b-5p	3918	hsa-miR-200c-3p	2316	hsa-miR-204-5p	1948
hsa-miR-199b-5p	3091	hsa-miR-200c-3p	23326	hsa-miR-204-5p	249
hsa-miR-199b-5p	2064	hsa-miR-204-5p	9586	hsa-miR-204-5p	50964
hsa-miR-200c-3p	10381	hsa-miR-204-5p	4211	hsa-miR-204-5p	2186
hsa-miR-200c-3p	10381	hsa-miR-204-5p	3206	hsa-miR-204-5p	5573
hsa-miR-200c-3p	648	hsa-miR-204-5p	596	hsa-miR-204-5p	65267
hsa-miR-200c-3p	4478	hsa-miR-204-5p	2296	hsa-miR-204-5p	6482
hsa-miR-200c-3p	8487	hsa-miR-204-5p	7048	hsa-miR-204-5p	64599
hsa-miR-200c-3p	8314	hsa-miR-204-5p	6591	hsa-miR-204-5p	6176
hsa-miR-200c-3p	93	hsa-miR-204-5p	7046	hsa-miR-204-5p	3320
hsa-miR-200c-3p	9839	hsa-miR-204-5p	81631	hsa-miR-204-5p	905
hsa-miR-200c-3p	29117	hsa-miR-204-5p	4212	hsa-miR-204-5p	10527
hsa-miR-200c-3p	6935	hsa-miR-204-5p	6615	hsa-miR-204-5p	100124696
hsa-miR-200c-3p	54206	hsa-miR-204-5p	25803	hsa-miR-204-5p	10813
hsa-miR-200c-3p	2033	hsa-miR-204-5p	7068	hsa-miR-204-5p	26061
hsa-miR-200c-3p	2335	hsa-miR-204-5p	1045	hsa-miR-204-5p	2023
hsa-miR-200c-3p	23414	hsa-miR-204-5p	8165	hsa-miR-204-5p	9097
hsa-miR-200c-3p	7329	hsa-miR-204-5p	1174	hsa-miR-204-5p	6046
hsa-miR-200c-3p	182	hsa-miR-204-5p	8905	hsa-miR-204-5p	554313
hsa-miR-200c-3p	5783	hsa-miR-204-5p	9411	hsa-miR-204-5p	7812
hsa-miR-200c-3p	4915	hsa-miR-204-5p	490	hsa-miR-204-5p	6122
hsa-miR-200c-3p	9134	hsa-miR-204-5p	599	hsa-miR-204-5p	84707
hsa-miR-200c-3p	331	hsa-miR-204-5p	329	hsa-miR-22-5p	29070
hsa-miR-200c-3p	596	hsa-miR-204-5p	1000	hsa-miR-22-5p	5901
hsa-miR-200c-3p	10516	hsa-miR-204-5p	50509	hsa-miR-22-5p	3178
hsa-miR-200c-3p	7422	hsa-miR-204-5p	2921	hsa-miR-22-5p	6134
hsa-miR-200c-3p	3551	hsa-miR-204-5p	9695	hsa-miR-22-5p	9774
hsa-miR-200c-3p	2321	hsa-miR-204-5p	7430	hsa-miR-22-5p	54471
hsa-miR-200c-3p	687	hsa-miR-204-5p	10160	hsa-miR-506-3p	2395
hsa-miR-200c-3p	29110	hsa-miR-204-5p	8321		

Tabella A.2: Geni target dei microRNA differenzialmente espressi.

PATHWAY	*p*-value espressione		*p*-value metilazione	
	Varianza	Media	Varianza	Media
Acute myeloid leukemia	0.01	0.01	0.01	0.00
Adrenergic signaling in cardiomyocytes	0.03	0.01	0.22	0.01
Alcoholism	0.00	0.01	0.31	0.00
Amoebiasis	0.00	0.00	0.13	0.01
Apoptosis	0.02	0.00	0.29	0.00
Bacterial invasion of epithelial cells	0.05	0.00	0.01	0.01
Butanoate metabolism	0.00	0.01	0.09	0.00
Chemokine signaling pathway	0.41	0.01	0.30	0.01
Chronic myeloid leukemia	0.03	0.00	0.11	0.00
Cocaine addiction	0.00	0.01	0.30	0.01
Cytokine-cytokine receptor interaction	0.02	0.00	0.06	0.01
Dilated cardiomyopathy	0.00	0.01	0.29	0.00
Epstein-Barr virus infection	0.04	0.01	0.13	0.00
Ether lipid metabolism	0.00	0.01	0.11	0.01
GABAergic synapse	0.01	0.00	0.12	0.01
Glioma	0.00	0.00	0.04	0.00
Glycerolipid metabolism	0.15	0.00	0.57	0.00
Glycerophospholipid metabolism	0.39	0.00	0.79	0.00
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	0.22	0.00	0.24	0.00
Hepatitis B	0.05	0.00	0.08	0.00
Hepatitis C	0.01	0.00	0.13	0.01
Herpes simplex infection	0.07	0.01	0.05	0.01
HIF-1 signaling pathway	0.01	0.01	0.11	0.00
HTLV-I infection	0.02	0.01	0.13	0.00
Linoleic acid metabolism	0.10	0.00	0.09	0.00
Measles	0.00	0.01	0.01	0.00
MicroRNAs in cancer	0.13	0.00	0.01	0.00
Natural killer cell mediated cytotoxicity	0.04	0.00	0.02	0.01
Neurotrophin signaling pathway	0.02	0.00	0.16	0.00
Osteoclast differentiation	0.01	0.00	0.05	0.01
p53 signaling pathway	0.44	0.00	0.28	0.00
Pancreatic cancer	0.01	0.00	0.12	0.01
Phototransduction	0.00	0.00	0.51	0.00
Propanoate metabolism	0.19	0.00	0.17	0.01
Prostate cancer	0.00	0.00	0.01	0.01
Ras signaling pathway	0.10	0.00	0.64	0.00
Regulation of actin cytoskeleton	0.01	0.00	0.11	0.00
Renal cell carcinoma	0.02	0.00	0.07	0.00
RIG-I-like receptor signaling pathway	0.10	0.00	0.11	0.00
Tuberculosis	0.00	0.01	0.09	0.01
Tyrosine metabolism	0.07	0.01	0.07	0.01
Vascular smooth muscle contraction	0.06	0.01	0.08	0.01
Wnt signaling pathway	0.01	0.01	0.11	0.00

Tabella A.3: *p*-value dei *pathway* significativi in media per l'espressione e la metilazione dei geni codificanti.

Bibliografia

[1] Anderson (2003) TW *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

[2] Benjamini Y. & Hochberg, Y. (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. Journal of the Royal Statistical Society. Series B, 57(1), pp. 289-300.

[3] Cancer Genome Atlas Research Network. (2011). *Integrated genomic analyses of ovarian carcinoma*. Nature, 474(7353), 609-615.

[4] Chiogna M, Massa MS, Risso D, Romualdi (2009). *A comparison on e ects of normalisations in the detection od di erentially espressed genes*. BMA Bionformatics 2009; 10:61.

[5] Fienberg, S. E. (1970). *An iterative procedure for estimation in contingency tables*. The Annals of Mathematical Statistics, 907-917.

[6] He, L., & Hannon, G. J. (2004). *MicroRNAs: small RNAs with a big role in gene regulation*. Nature Reviews Genetics, 5(7), 522-531.

[7] Khatri, P., Sirota, M., & Butte, A. J. (2012). *Ten years of pathway analysis: current approaches and outstanding challenges*. PLoS computational biology, 8(2), e1002375.

[8] Lauritzen SL (1996). *Graphical models*. Clarendon Press, Oxford.

[9] Martini, P., Sales, G., Massa, M. S., Chiogna, M., & Romualdi, C. (2013). *Along signal paths: an empirical gene set approach exploiting pathway topology*. Nucleic acids research, 41(1), e19-e19.

[10] Massa, M. S., Chiogna, M., & Romualdi, C. (2010). *Gene set analysis exploiting the topology of a pathway*. BMC Systems Biology, 4(1), 121.

[11] Mitrea C., Taghavi Z., Bokanizad B., Hanoudi S., Tagett R., Donato M., Voichita C. & Draghici S. (2013). *Methods and approaches in the topology-based analysis of biological pathways*. Frontiers in physiology, 4.

[12] Robertson, K. D. (2005). *DNA methylation and human disease*. Nature Reviews Genetics, 6(8), 597-610.

[13] Sales, G., Calura, E., Cavalieri, D., & Romualdi, C. (2012). *graphite-a Bioconductor package to convert pathway topology to gene network*. BMC bioinformatics, 13(1), 20.

[14] Sales, G., Calura, E., & Romualdi, C. (2011). *GRAPH Interaction from pathway Topological Environment*.

[15] Schäfer J., Strimmer K. (2005) *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*. Statistical applications in genetics and molecular biology, Volume 4, Issue 1, Article 32.

[16] Smyth, G. K. (2004). *Statistical Applications in Genetics and Molecular Biology*. Volume 3, Issue 1, Article 3.

[17] Storey, J. & Tibshirani, R. (2001) *Estimating false discovery rates under dependence, with applications to DNA microarrays*. Stanford: Stanford University, Department of Statistics; Report No.: Technical Report 200128, pp. 7-9.

[18] Storey, J. & Tibshirani, R. (2003). *Statistical significance for genomewide studies*. Proceedings of the National Academy of Sciences, 100(16), pp. 9440-9445.