Università degli Studi di Padova

Corso di Laurea in Scienze Statistiche

Analisi di pathway biologici: studio e confronto di tecniche topologiche

Relatore: Prof. Dott. Chiara Romualdi

Dipartimento di Biologia

Laurendo: Anna Turrin

Anno Accademico 2011/2012

Indice

1 Introduzione

La bioinformatica, nata alla fine degli anni 70, è una scienza che si sta rapidamente evolvendo. Nata come branca della biologia, è una disciplina altamente interdisciplinare che applica a problemi di natura biologica numerose tecniche e concetti derivanti dall'informatica, statistica, matematica, fisica, chimica e biochimica. Il *National Center for Biointecnology Information* (NCBI) definisce la bioinformatica come "*la scienza nella quale biologia, informatica e tecnologia dell'informazione si uniscono in un'unica disciplina*". [1]

Una delle possibili analisi di dati che si possono fare attraverso la bioinformatica è l'analisi dei dati di esperimenti di *microarray*. Un *microarray di DNA* (o matrici ad alta densità) è costituito da un insieme di microscopiche sonde di DNA, dette *array*, attaccate a un vetrino. Questi *array* sono usati per esaminare il profilo d'espressione di un gene o per identificare la presenza di un gene, o di una breve sequenza, all'interno di una miscela di migliaia di geni.

[1] http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

1.1 Nozioni di base della biologia molecolare

Il DNA venne inizialmente isolato dal biologo svizzero Friedrich Miescher nel 1869, ma solo nel 1953 James Watson e Francis Crick presentano il primo modello accurato della struttura di DNA, quello a doppia elica. Nel 1957 sempre Crick propose il dogma centrale della biologia molecolare che fissa le relazioni tra DNA, RNA e proteine, ma solo successivamente egli dimostrò come il codice genetico fosse basato su triplette di basi non sovrapposte, permettendo così a Har Gobind Khorana, Robert Holley e Marshall Warren Nirenberg di decifrarlo. Queste scoperte sono state la base per lo sviluppo della nuova biologia molecolare.

Il DNA, come è stato detto, ha una struttura a doppia elica, simile a una scala a pioli disposta a spirale. Lo scheletro di questa "scala" è composto da zucchero e un gruppo fosfato, mentre i pioli sono composti da quattro basi azotate (adenina (A), citosina (C), guanina (G) e timina (T)). Ogni base presente su un filamento si lega in modo univoco a una base del filamento opposto:

- A appaia con T
- G appaia con C

La disposizione in sequenza di queste 4 basi costituisce l'informazione genetica, leggibile attraverso il codice genitico, che ne permette la traduzione in amminoacidi. Il gene corrisponde a una porzione di codice genitico

all'interno della sequenza di DNA. Il processo di traduzione (figura 1.1) del DNA in proteine avviene in due fasi:

- La trascrizione;
- La traduzione;

L'informazione contenuta in un filamento di DNA viene trascritta in filamento di RNA messaggero (mRNA), che successivamente esce dal nucleo, si sposta sui ribosomi dove interviene l'RNA di trasporto (tRNA), costituito da una tripletta di basi azotate specifica per l'amminoacido che trasporta. In questo modo gli amminoacidi vengono allineati secondo la sequenza iniziale di DNA e formano la catena peptidica, ovvero una proteina. Durante il processo di traduzione possono avvenire dei cambiamenti, chiamate mutazioni, della sequenza delle basi, come l'aggiunta, la sostituzione o l'eliminazione di un nucleotide. Questo porta a delle mutazioni nei geni che codificano alcune proteine, che possono quindi diventare inattive oppure essere addirittura mancanti. Questo può essere la causa dell'insorgenza di molte malattie genetiche.

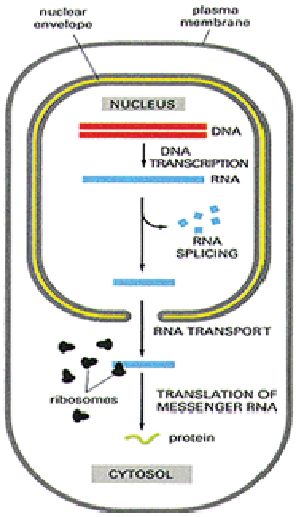

Figura 1.1: Il processo di traduzione del DNA

1.2 La tecnologia di *Affymetrix*

Come già detto in precedenza la bioinformatica ci consente di fare analisi dei dati di esperimenti di *microarray*. La tecnologia dei *microarray* ci permette di misurare l'espressione genica, in differenti condizioni, su migliaia di geni simultaneamente. Esistono vari tipi di microarray in commercio, ma nel presente elaborato si farà riferimento solo alla tecnologia a singolo canale prodotta da *Affymetrix*. Gli esperimenti a singolo canale sono caratterizzati da un solo campione per *chip* (figura 1.2) e forniscono una misura diretta ed indipendente del livello di espressione genica in un campione.

La tecnologia dei *microarray* a DNA si basa sulla capacità di ibridazione degli acidi nucleici, ovvero due filamenti di DNA riescono a ibridizzare se sono tra di loro complementari. La tecnologia considerata in questo elaborato utilizza sonde (o *probe*) lunghe circa 25 paia di basi che vengono bloccate su una superficie di vetro (figura 1.3). In ogni vetrino sono presenti alcune centinaia di migliaia di filamenti, ciascuno dei quali presente in milioni di copie. Ogni gene o sequenza genomica di interesse è rappresentato da un *probeset* formato da 11 a 20 coppie di probe (*probe pair*). Ogni coppia è formata da un *perfect match (PM) probe* e da un *mismatch (MM) probe* creato con la tredicesima base diversa. Il livello di espressione di un gene, in questo tipo di tecnologia, è ottenuto dall'integrazione di tutti i dati del *probeset*.

Figura 1.2: Il GeneChip

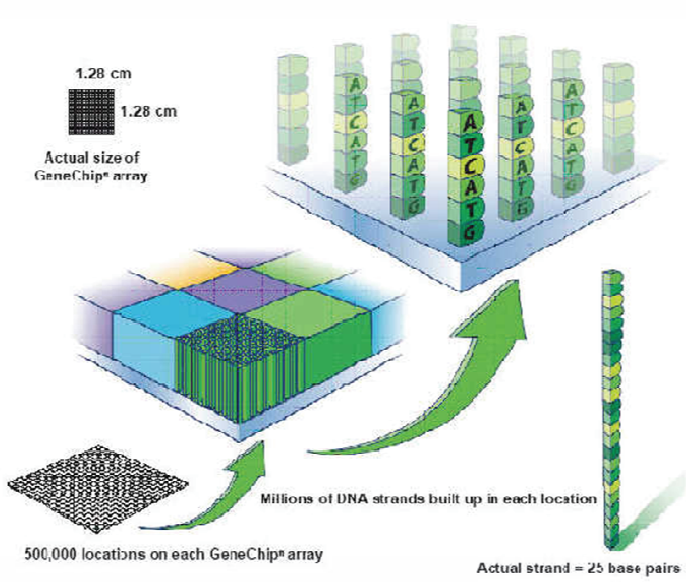

Figura 1.3: I probe nel GeneChip array

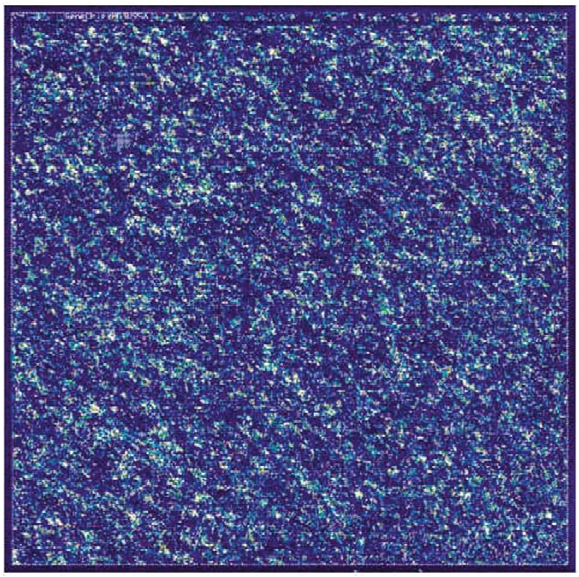

Figura 1.4: Scansione del GeneChip con scanner laser

1.2.1 La matrice dei dati

Una volta creato il *GeneChip* tramite ibridizzazione e scansionato con l'utilizzo di uno scanner laser, otteniamo l'immagine di tutti gli spot che sono stati creati (figura 1.4). Una volta che è stata identificata la posizione di questi spot si può passare al calcolo delle intensità di ogni spot e al calcolo del *background* locale. Dall'immagine scansionata, quindi, si passa alla matrice di intensità, dove nel senso delle righe abbbiamo i diversi spots e nel senso delle colonne abbiamo i diversi valori di intensità per ogni spot. Nel caso particolare della tecnologia *Affymetrix* i dati grezzi non sono presentati come una matrice di dati, ma sono presentati in file .CEL, il cui formato è un file di testo ASCII, da cui successivamente potremo ottenere la matrice dei dati (tabella 1.1). Questo dataset si presenta come una matrice con n righe e p colonne, dove nel senso delle righe si trovano i vari geni, le unità di interesse, e nel senso delle colonne si trovano i vari esperimenti.

	Exp.1	Exp.2	$\cdots$	Exp.p
Gene 1	x_{11}	x_{12}	$\cdots$	x_{1p}
Gene 2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Gene n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Tabella 1.1: **Esempio di matrice dei dati**

1.2.2 La normalizzazione

Uno dei punti critici dei dati di *microarray* sta nella normalizzazione. Una matrice di dati contiene un numero molto alto di geni, a fronte di un numero molto più limitato di esperimenti. Si può comprendere dunque come anche un piccolo errore sistematico possa creare delle distorsioni significative durante la fase di inferenza. La normalizzazione dei dati avviene, quindi, in una fase preliminare all'inferenza. La normalizzazione tra *array*, eseguita principalmente negli *array* a singolo canale, permette di riscontrare delle differenze di intensità dovute a diverse potenze dei laser o ai parametri di scansione. In questo caso lo scopo di questa normalizzazione è quello di rendere confrontabili i valori di espressione di tutti gli esperimenti considerati.
La maggior parte delle procedure di normalizzazione per la riduzione degli errori sistematici si basa su tre assunzioni:

- Pochi geni differenzialmente espressi nelle due condizioni sperimentali;
- Il numero di geni sovra espressi sia simile al numero di geni sotto espressi;
- La differenziale espressione non dipenda dalla media del segnale.

1.3 I geni differenzialmente espressi (DEG)

Negli esperimenti di *microarray* un risultato importante dell'analisi inferenziale è quello dell'identificazione dei geni differenzialmente espressi

(DEG). Vengono classificati con questo nome i geni che mostrano una significativa differenza nei livelli di espressione in due o più gruppi di campioni biologici oggetto di studio. In letteratura si trovano molti test che possono essere utilizzati per l'identificazione di tali geni. Si rimanda al Capitolo 2 per il dettaglio delle tecniche statistiche usate in questo elaborato. Va ricordato, inoltre, che i test per identificare i geni DEG vanno condotti su ogni singolo gene. In altre parole va effettuato un test per ogni riga della matrice dei dati e per ogni gene si avrà un livello di significatività. Il numero dei geni differenzialmente espressi dipenderà, quindi, da un livello soglia (*cut-off*) oltre il quale si rifiuta l'ipotesi di uguaglianza dei valori di espressione nei due, o più, gruppi. Come è stato precedentemente detto il numero di geni presenti in ogni esperimento è dell'ordine di decine di migliaia, e questo porta inevitabilmente a un problema di confronti multipli. E' noto che l'utilizzo di molti confronti multipli comporta un incremento dell'errore di I tipo, in quanto a livello globale, la probabilità di rifiutare erroneamente un'ipotesi nulla aumenta con l'aumentare dei confronti. Molto spesso si utilizzano procedure che controllano il *Family Wise Error Rate*[16] (FWER), definito come la probabilità di avere almeno un falso positivo fra tutti i test fatti, come la correzione di Bonferroni o la correzione di Holm-Sidak. Un'altra tecnica che è molto spesso utilizzata per controllare il valore di soglia è il *False Discovery Rate* (FDR). Questo metodo è stato proposto per la prima volta nel 1995 da Benjamini e Hochberg[16] e risulta essere un buon compromesso tra l'esigenza di tenere sotto controllo il rischio di commettere errori di I tipo, che aumenta con l'aumentare dei confronti,

e la necessità di evitare un'eccessiva riduzione della potenza del test, che diminuisce quanto più si abbassa la probabilità di soglia per l'errore di primo tipo. Il *False Discovery Rate* è definito come il numero atteso di falsi positivi, ovvero è il rapporto tra il numero di geni che si dichiarano significativi, o differenzialmente espressi, quando l'ipotesi nulla è vera e il numero totale di geni differenziamente espressi. Da questo ne consegue che tenere una soglia di FDR bassa equivale a tenere bassa la probabilità di ottenere falsi positivi.

1.4 Lo studio topologico dei pathway

Una volta svolta l'analisi differenziale si ottiene una lista di qualche centinaia di geni differenzialmente espressi. La fase di interpretazione dei risultati, quindi, risulta molto complessa soprattutto nel caso in cui il problema biologico studiato sia poco noto. Si cerca di risolvere questo problema utilizzando le annotazioni funzionali dei geni usando la *Gene Ontology* oppure il database di pathway *KEGG* (parte della *Kyoto Encyclopedia of Genes and Genoms*) con i quali si cerca di raggruppare i geni in categorie. Con questo tipo di analisi si cerca di ricostruire il fitto scambio di informazioni fra i geni che avviene in una certa condizione sperimentale. Il *database* KEGG, che verrà utilizzatato anche nelle analisi successive, contiene informazioni sui percorsi di co-regolazione che sono alla base della trasmissione del segnale genetico. In essa sono presenti le mappe (chiamate *pathway*) che visualizzano in modo grafico i diffrenti livelli di interazione fra i geni. Queste informazioni ci consentono di

realizzare le analisi di *Gene Set*, o analisi funzionali, ovvero un tipo di analisi che mira ad identificare nella lista di geni DEG particolari informazioni biologiche. Questo genere di analisi si dividono essenzialmente in due tipi:

- Quelle che sovrappongono semplicemente la lista di geni DEG alle mappe di co-regolazione, *Over Representation Analysis* (ORA) ;
- Quelle che cercano di associare un parametro statistico ad un *pathway* che sia esplicativo non solo delle differenze di espressione, ma anche della loro posizione topologica.

Un *pathway* metabolico (figura 1.5) è un diagramma nel quale sono rappresentate le reazioni chimiche coinvolte in uno o più processi cellulari. I singoli passi del *pathway* sono reazioni, catalizzate nella maggior parte dei casi da enzimi, che agiscono su una molecola di partenza, chiamata substrato, trasformandola in un prodotto, il quale sarà utilizzato a sua volta come substrato al passo successivo della reazione chimica. Il *pathway* in particolare può essere studiato da un punto di vista strutturale attraverso un modello basato sui grafi. Utilizzare un grafo per modellare una rete metabolica (o *network*) significa scegliere quali entità biologiche associare a nodi e archi. Un *network* metabolico (figura 1.6), quindi, è un insieme di nodi costituiti da prodotti genici, in particolare proteine, ma anche RNA e complessi, relazionati tra di loro tramite archi orientati a seconda del tipo di interazioni funzionali da cui sono legate.

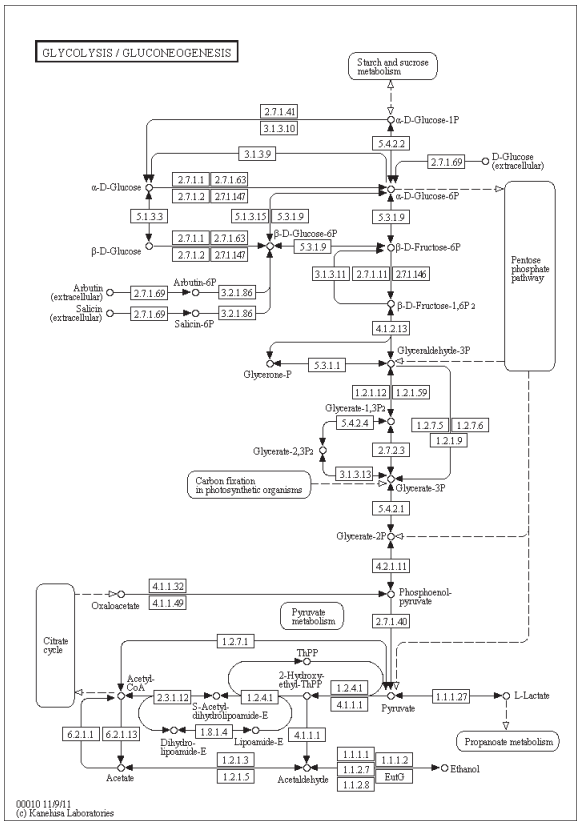

Figura 1.5: Esempio di pathway metabolico

(http://www.genome.jp/kegg/pathway/map/map00010.html)

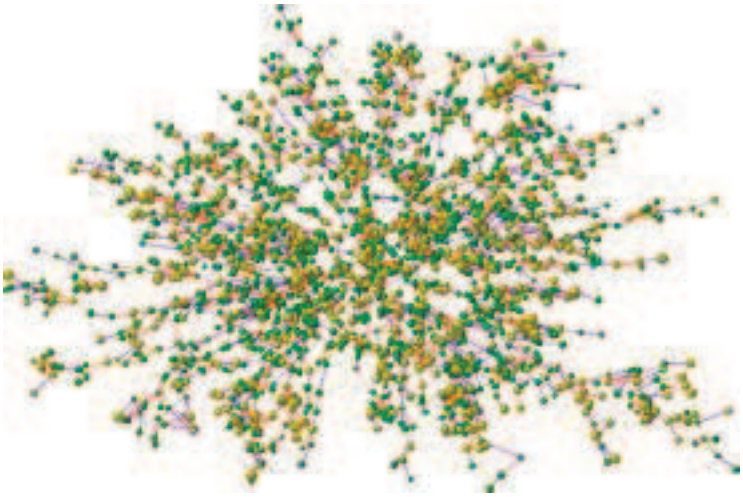

Figura 1.6: Esempio di network

1.5 Scopo della tesi

Come detto precedentemente il *pathway*, che fornisce importanti informazioni sulla correlazione esistente tra i geni, è rappresentato come un grafo i cui nodi corrispondono molto spesso a prodotti genici. Questi possono essere divisi in due tipi:

- complessi proteici (gruppi AND), che possono essere espansi in *cliques* in cui tutte le proteine sono collegate tra di loro;
- gruppi con elementi tra di loro sostituibili, come ad esempio le famiglie geniche o geni che hanno funzioni biochimiche simili (gruppi OR).

Il fatto di tener conto della presenza di questi diversi tipi di prodotti genici, porta ad avere delle stime più precise nell'identificazione di sottogruppi genici sregolati e riesce a evidenziare quei segnali che, all'interno del *pathway*, sono più coinvolti nella sua sregolatezza. Il *pathway*, in questo modo, però diventa molto più complesso non solo da un punto di vista topologico, ma anche computazionale. L'interesse di questo elaborato è quello di cercare e valutare un metodo che consenta, partendo da un *pathway* in cui sia i gruppi AND che OR sono stati espansi, di semplificare la struttura del *pathway* per rendere quindi anche più veloce l'analisi. Il *pathway* considerato è quello di Chronic Myeloid Leukemia. La leucemia mieloidale cronica è una delle prime malattie per cui si è individuata una specifica anormalità cromosomica tra le sue cause. Nello specifico la traslocazione del gene ABL dal cromosoma 9 al cromosoma

22 con conseguente formazione di un gene chimera.

Si sono considerate:

- Il *dataset* relativo al *pathway* "Chronic Myeloid Leukemia" in cui i gruppi AND e i gruppi OR sono stati espansi (nel seguito "Iniziale");
- Il *dataset* in cui i gruppi AND e OR sono rappresentati dalla media dei geni che li compongono ("Medie");
- Il *dataset* in cui i gruppi OR sono rappresentati dal gene che mostra una maggiore differenziale espressione in media nelle due condizioni sperimentali, mentre i gruppi AND vengono rappresentari sempre attraverso la media semplice ("Medie e Massima Differenziale Espressione");
- Il *dataset* in cui sia i gruppi AND che OR sono rappresentati dal gene che mostra una maggiore differenziale espressione in media nelle due condizioni sperimentali ("Massima Differenziale Espressione");
- Il *dataset* in cui entrambe le tipologie di gruppi vengono espresse secondo il metodo delle componeti principali ("Componente Principale");

Su questi cinque *dataset* sono state svolte le stesse analisi, ovvero identificazione dei geni differenzialmente espressi e analisi topologiche sui *pathway* con TopologyGSA e SPIA. Sono stati confrontati poi i risultati ottenuti dalle quattro matrici semplificate con quelli ottenuti dalla matri-

ce dei dati iniziale. Uno schema generale di come è stata svolta l'analisi è indicato in figura 1.5.

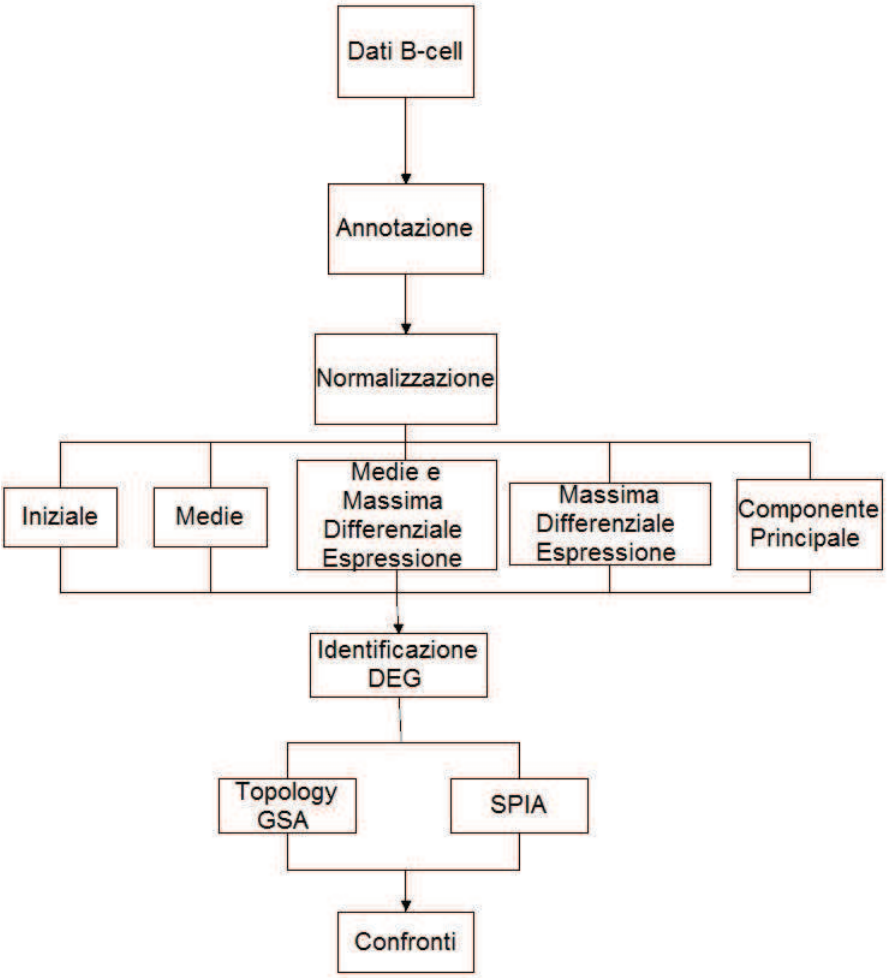

Figura 1.5: Struttura dell'analisi svolta

2 Tecniche statistiche per l'identificazione dei DEG

2.1 Introduzione

Come già accennato, l'identificazione di geni differenzialmente espressi è uno dei risultati di maggior interesse quando si parla di esperimenti di *microarray*. Questa fase inferenziale, che avviene successivamente alla fase di normalizzazione, ha lo scopo, dunque, di individuare i geni che hanno un valore di espressione significativamente diverso tra due o più condizioni sperimentali. L'ipotesi nulla, del sistema d'ipotesi a cui ci si riferisce, esprime l'uguaglianza dell'espressione del gene g nelle due condizioni sperimentali e l'ipotesi alternativa ne rappresenta la sua negazione. Formalizzando si ha:

$$H_{0,g} : \mu_x(g) = \mu_y(g) \tag{2.1}$$

$$H_{1,g} : \mu_x(g) \neq \mu_y(g)$$

con $g = 1, ..., n$, dove $\mu_x(g)$ è il valore atteso dell'espressione del gene g nella prima condizione e $\mu_y(g)$ è il valore atteso dell'espressione del gene g nella seconda condizione. In letteratura esistono vari tipi di test utilizzati con lo scopo di identificare i geni differenzialmente espressi, che adottano sia un approccio parametrico che non parametrico, ma in questo elaborato verrà introdotta solo la parte relativa ai modelli lineari per dati di *microarray* (in quanto verranno utilizzati poi nelle analisi successive). Va ricordato che i problemi che si riscontrano nei criteri utilizzati nell'inferenza classica sono principalmente due:

- Il controllo dell'errore di primo tipo (α). Sono stati proposti alcuni metodi di aggiustamento dei valori-p stimati che controllano il *family wise error rate* o, in alternativa, il *false discovery rate*.
- La stima della varianza gene-specifica (σ_g^2).

Inoltre, quando si parla di applicare l'inferenza classica ai dati di *microarray* bisogna tener presente che il sistema d'ipotesi (2.1) deve essere testato su ogni singolo gene della piattaforma (che contiene centinaia di migliaia di geni), e non è detto che siano rispettate le assunzioni di normalità della distribuzione, di omoschedasticità e di indipendenza.

2.2 Il modello lineare per dati di microarray

2.2.1 La matrice del disegno

I modelli lineari vengono utilizzati per fare inferenza sia per dati che derivano da tecnologie a singolo canale, sia per dati che derivano da

tecnologie a doppio canale. In generale si assume di avere un *dataset* con p *microarray* dai quali si ottiene un vettore (vettore risposta) contenente i valori di espressione, per ogni singolo gene. Per il g - esimo gene si avrà quindi il vettore:

$$\mathbf{y_g} = \begin{bmatrix} y_{g1} \\ y_{g2} \\ \cdots \\ y_{gp} \end{bmatrix} \tag{2.2}$$

Il valore di espressione, generalmente, è presentato in termini di logaritmo in base 2 del rapporto tra le intensità delle due condizioni sperimentali per le tecnologie a doppio canale e in termini di logaritmo in base 2 delle intensità nel caso di tecnologie a singolo canale. Il modello che viene dunque assunto, per ogni singolo gene, è del tipo:

$$\mathbf{y_g} = \mathbf{X}\alpha_\mathbf{g} + \epsilon_\mathbf{g} \tag{2.3}$$

dove $\mathbf{X}$ rappresenta la matrice del disegno, α_g è il vettore dei parametri e ϵ_g è un fattore d'errore non necessariamente normale. La matrice del disegno nel caso di tecnologie a singolo canale [5] è uguale a quella utilizzata nei modelli lineari classici, ovvero, supponendo di avere i array per il primo gruppo e j array per il secondo gruppo $(i + j = p)$, una possibile matrice del disegno sarà del tipo

$$\mathbf{X_{(p\times 2)}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \dots & \dots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} i - ma\, posizione \tag{2.4}$$

2.2.2 Il t-test

Riprendendo il modello (2.3) del paragrafo precedente si assume che

$$E(\mathbf{y}_g) = X\alpha_g \tag{2.5}$$

$$var(\mathbf{y}_g) = W_g\sigma_g^2 \tag{2.6}$$

dove W_g è una matrice di pesi, nota, definita non negativa. Si assume, inoltre, che alcuni contrasti siano di interesse biologico, e questi possono essere definiti come:

$$\beta_g = C^T\alpha_g \tag{2.7}$$

Infine, si assume che sia di interesse testare se il singolo valore del contrasto β_{gj} sia uguale a zero, ovvero $\beta_{gj} = 0$. Ne consegue, quindi, che il sistema di ipotesi sia del tipo:

$$H_0 : \beta_{gj} = 0 \tag{2.8}$$

$$H_1 : \beta_{gj} \neq 0$$

Sul vettore delle risposte viene dunque stimato il modello lineare in modo da ottenere, per ogni gene, una stima dei seguenti coefficienti:

- $\hat{\alpha_g}$ per α_g;
- s_g^2 stima per σ_g^2

dai quali si ottiene poi una stima della matrice di varianze e covarianze

$$var(\hat{\alpha}_g) = V_g s_g^2 \tag{2.9}$$

dove V_g è una matrice definita positiva, che non dipende da s_g^2. Una volta ottenute queste stime si ottiene una stima per i coefficienti β_g che definiscono i contrasti

$$\hat{\beta}_g = C^T \hat{\alpha}_g \tag{2.10}$$

la cui matrice di varianze e covarianze stimata è

$$var(\hat{\beta}_g) = C^T V_g C s_g^2 \tag{2.11}$$

Il modello non è necessariamente assunto normale, e quindi non va necessariamente usata una stima ai minimi quadrati. Ad ogni modo, si assume che la stima dei contrasti sia approssimativamente normale con media β_g

e matrice di varianze e covarianze $C^T V_g C s_g^2$ e che la varianza residua s_g^2 segua approssimativamente una distribuzione χ^2 scalata. Definendo con v_{gj} l'elemento diagonale j-esimo della matrice $C^T V_g C$ le assunzioni sulle distribuzioni possono essere riassunte come segue:

$$\hat{\beta}_{gj} \sim N(\beta_{gj}, v_{gj}\sigma_g^2) \tag{2.12}$$

$$s_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2 \tag{2.13}$$

dove d_g sono i gradi di libertà residui del modello lineare per il gene g. Sotto queste assunzioni si può definire la statistica test *t ordinaria*

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}} \sim t_{dg} \tag{2.14}$$

che segue approssimativamente la distribuzione *t-Student* con dg gradi di libertà. Diventa, però, complesso dal punto di vista computazione applicare questa statistica test a dati di *microarray*, in quanto, come detto precedentemente, in ogni piattaforma sono presenti centinaia di migliaia di geni. Per questo motivo accade di frequente che alcuni geni presentino somme di quadrati molto piccole, o molto grandi, producendo di conseguenza valori del test, rispettivamente, molto grandi, o molto piccoli, pur essendo espressi in quantità scarse. Per evitare il problema che questi geni generino degli errori di identificazione sono state proposte delle statistiche alternative, che consentono agli stimatori delle varianze di ridurre l'effetto dei valori esageratamente grandi e esageratamente piccoli.

Sono stati introdotti quindi degli stimatori *shrinkage*, che vengono usati in statistiche chiamate *t-moderate*, con lo scopo di ridurre la distorsione delle stime. Alcune soluzioni proposte per questo genere di stimatori sono influenzate dalla teoria bayesiana e dalla teoria bayesiana empirica. I metodi bayesiani si adattano bene allo studio di dati di *microarray*, dato che sono usati solitamente nei problemi di inferenza multidimensionale. La più grande differenza tra i metodi bayesiani e quelli dell'inferenza classica sta nel fatto che quest'ultima applica l'inferenza separatamente su ogni singolo gene, mentre i metodi bayesiani e bayesiani empirici utilizzano l'informazione proveniente da tutti i geni, che viene riassunta nella formulazione di distribuzioni a priori per alcuni parametri. Questi parametri vengono poi combinati a livello di medie e di deviazioni standard a livello dei geni.

2.2.3 Il test Bayesiano Empirico eBayes

Come già accennato, quando si fa inferenza negli esperimenti di *microarray*, è preferibile, considerare la struttura parallela dei dati per la quale lo stesso modello viene stimato su ogni singolo gene che compone la piattaforma in esame. E' appunto con questo scopo di semplificare la fase di inferenza che si utilizza un modello gerarchico. Il punto focale in questo tipo di modello è quello di cercare di capire come i coefficienti ignoti β_{gj} e σ_g^2 variano tra i geni, e questo viene fatto assumendo delle distribuzioni a priori per entrambi i parametri. Nello specifico si assumono come distribuzioni a priori:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2 \tag{2.15}$$

che descrive il modo in cui ci si aspetta che la varianza vari tra i geni, dove s_0^2 è l'iperparametro della distribuzione per σ_g^2 e d_0 indica i gradi di libertà.

$$P(\beta_{gj} \neq 0) = p_j \tag{2.16}$$

per ogni j si assume che β_{gj} sia diverso da 0 con una probabilità nota, ovvero p_j è la vera proporzione di geni differenzialmente espressi che ci si attende.

$$\beta_{gj} \mid \sigma_g^2,\, \beta_{gj} \neq 0 \sim N(0,\, v_{0j}\sigma_g^2) \tag{2.17}$$

che descrive la distribuzione dei *log-fold changes* (ovvero il logaritmo del rapporto tra le intensità del controllo e del trattato) per i geni che sono differenzialmente espressi. Date le assunzioni del modello gerarchico si definisce, poi, la media a posteriori di $\sigma_g^{-2} \mid s_g^2$, ovvero $\tilde{s}_g^{-2}$ dove

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \tag{2.18}$$

Il valore a posteriori avvicina le varianze osservate ai valori a priori e viene detto *stimatore shrinkage*. A questo punto si può definire la statistica *test t moderata*

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} \tag{2.19}$$

In letteratura [1] è stato dimostrato che $\tilde{t}$ e s_g^2 sono indipendenti in distribuzione, e in particolare, che la statistica t moderata segue una distribuzione t di Student con $d_g + d_0$ gradi di libertà, sotto l'ipotesi nulla $H_0 : \beta_{gj} = 0$. I gradi di libertà aggiuntivi rispetto alla statistica test t ordinaria, ovvero d_0, riflettono l'informazione aggiuntiva che si ottiene dall'insieme dei geni, sulla base del modello gerarchico qui sopra descritto. Un approccio completamente bayesiano consente di scegliere i parametri s_0 e d_0, in questo caso, invece, siamo di fronte a un approccio bayesiano empirico per cui i parametri sono stimati dai dati. In particolare s_0 e d_0 sono stimati uguagliando i primi due momenti attesi con quelli empirici della variabile $log(s_g^2)$. Una volta ottenute le distribuzioni $\tilde{t}_{gj}$ e s_g^2 si calcola facilmente il valore a posteriori che un gene sia differenzialmente espresso in relazione al contrasto β_{gj},

$$O_{gj} = \frac{p(\beta_{gj} \neq 0 \mid \tilde{t}_{gj},\, s_g^2)}{p(\beta_{gj} = 0 \mid \tilde{t}_{gj},\, s_g^2)} = \frac{p(\beta_{gj} \neq 0,\, \tilde{t}_{gj},\, s_g^2)}{p(\beta_{gj} = 0,\, \tilde{t}_{gj},\, s_g^2)} = \frac{p_j}{1 - p_j} \frac{p(\tilde{t}_{gj} \mid \beta_{gj} \neq 0)}{p(\tilde{t}_{gj} \mid \beta_{gj} = 0)} \tag{2.20}$$

Una trasformazione dell'ODD a posteriori è data dalla statistica B

$$B_{gj} = logO_{gj} \tag{2.21}$$

Un valore della statistica $B = 0$, rappresenta il caso di maggiore incertezza in quanto, derivando da un $ODD = 1$, rappresenta il caso in cui ho l'uguaglianza tra le due probabilità p e $1 - p$.

Fissare un valore soglia per la statistica B, per l'identificazione dei geni differenzialmente espressi, può non essere una buona idea in quanto

questa statistica si basa sul parametro p, sul quale sono state fatte delle ipotesi a priori. Di conseguenza questa statistica test viene usata prevalentemente come metodo di ordinamento dei geni sulla base dell'espressione differenziale. Un metodo alternativo può essere quello di usare il test t-moderato con correzione del *false discovery rate*, che viene fornito, assieme alla statistica B, come output della funzione `eBayes()` nelle analisi svolte con il software LIMMA. [1]

[1] E' possibile scaricare il pacchetto all'indirizzo web: http://bioconductor.org/packages/release/bioc/html/limma.html

3 Analisi topologiche sui pathway

3.1 Introduzione

Come è stato già accennato nel capitolo introduttivo di questo elaborato, recentemente gli studi su dati di *microarray* hanno generato un grande interesse da parte degli studiosi per quanto riguarda la *Gene Set Analysis*. Lo scopo della *Gene Set Analysis* è quello di individuare gruppi di geni correlati tra loro da un punto di vista funzionale. Descrivere ed estrarre informazioni da un *pathway* è lo scopo principale dei ricercatori delle varie discipline che sono coinvolte in questo genere di analisi. L'analisi topologica di un pathway consiste principalmente nello studio della sua struttura e delle relazioni che intercorrono tra le entità che lo compongono. Nonostante la topologia sia un'informazione preziosa, molti test utilizzati per la *Gene Set Analysis* non considerano le proprietà topologiche del *pathway*, ma considerano il *pathway* solo come semplice lista di geni. Solo recentemente sono stati proposti alcuni metodi che considerano entrambe le informazioni: l'appartenenza dei geni ad un *pa-*

thway e le loro interazioni. Il primo che verrà preso in considerazione è chiamato *Impact Analysis*, SPIA [6]. Questo tipo di analisi tiene in considerazione, oltre alle tecniche statistiche classiche, anche importanti aspetti biologici, come l'entità della variazione di ciascun gene, la loro tipologia e posizione all'interno di un dato *pathway* e le interazioni che li legano. SPIA, in particolare, da una maggiore importanza al *pathway* se i geni diffrenzialmente espressi sono collocati nella parte iniziale in quanto sono quei geni che poi influenzano a catena tutti gli altri geni che lo compongono. Un approccio alternativo è proposto anche da Massa et al. [7] la quale, però, basa il test sulla struttura di correlazione. In particolare, in questo approccio viene usata la teoria dei grafi per sviluppare il *pathway* complessivo in *cliques* (componenti connesse) più piccole per permettere, quindi, di analizzare con maggior dettaglio piccole porzioni dell'intero *pathway*.

3.1.1 Definizione del *pathway*

In questo elaborato si è considerato il pathway "*Chronic Myeloid Leukemia*" (figura 3.1) preso dal database di *KEGG Pathways* [1]. *KEGG Pathways* contiene mappe che descrivono le vie metaboliche e di segnale della cellula, ponendo anche attenzione alle variazioni che queste hanno tra diversi organismi viventi. I *pathway* sono disponibili in formato macchina (generalmente varianti di linguaggio xml) e formati grafici (interpretazioni del linguaggio macchina). *KEGG Pathway* utilizza come formato

[1] http://www.genome.jp/kegg/pathway.html

macchina un formato KGML (*KEGG Markup Language*)[2] che permette l'interpretazione automatica e la visualizzazione grafica del *pathway* e offre la possibilità di eseguire analisi computazionali e di modellazione di network proteici e chimici. All'interno di questo database, inoltre, per ogni *pathway*, si possono ottenere informazioni di tipo genomico, chimico e fenotipico su ognuna delle entità che lo compongono. Ogni *pathway* è composto da due diversi tipi di forme geometriche:

- I rettangoli;
- I cerchi;

I primi, nella versione più piccola, rappresentano i prodotti genici, nella maggior parte proteine, ma anche RNA e complessi. I rettangoli più grandi, invece, rappresentano dei link verso altri *pathway*, con cui quello in esame interagisce. I cerchi, infine, rappresentano composti chimici, chiamati *compound*, molto frequenti all'interno delle mappe (come ad esempio l'idrogeno, H_2O, ATP ...), che si comportano come un ponte tra due elementi. I rettangoli sono tra di loro collegati mediante archi che ne rappresentano le interazioni funzionali. Questi possono essere di diverse tipologie, come, per esempio, essere direzionati o non direzionati, tratteggiati direzionati o tratteggiati troncati; possono contenere anche informazioni riguardanti la fosforilazione (+p), la defosforilazione (-p) e altre reazioni chimiche.

[2] http://www.kegg.jp/kegg/xml/docs/

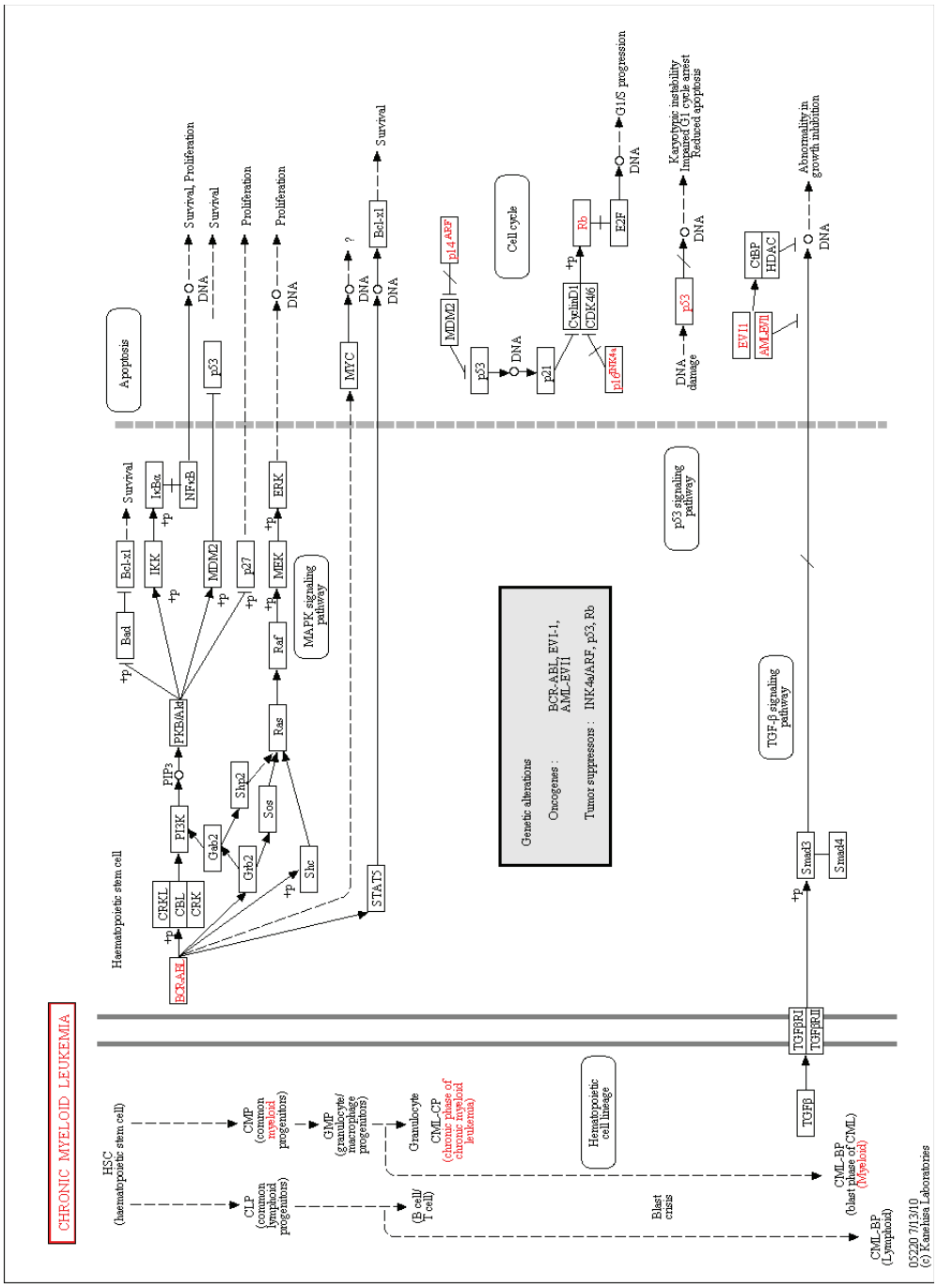

Figura 3.1: Il pathway **"Chronic Myeloid Leukemia"**

3.1.2 Conversione del *pathway* in *network*

Per svolgere delle analisi su un *pathway* bisogna prima di tutto trasformarlo in un *network*. Il *network* è una trasformazione grafica del *pathway*, che nel caso di KEGG viene resa possibile grazie al formato KGML con cui vengono tradotte le mappe. Un aspetto importante da notare in un *pathway*, una volta che è stato convertito in un *network* è che a un nodo del *pathway* corrispondono più prodotti genici, e quindi più nodi nel *network*. Questi possono essere divisi in due tipologie[8]:

- Gruppi AND (figura 3.2), ovvero complessi proteici, che possono essere espansi in *cliques*, dove ogni proteina è connessa con le altre;

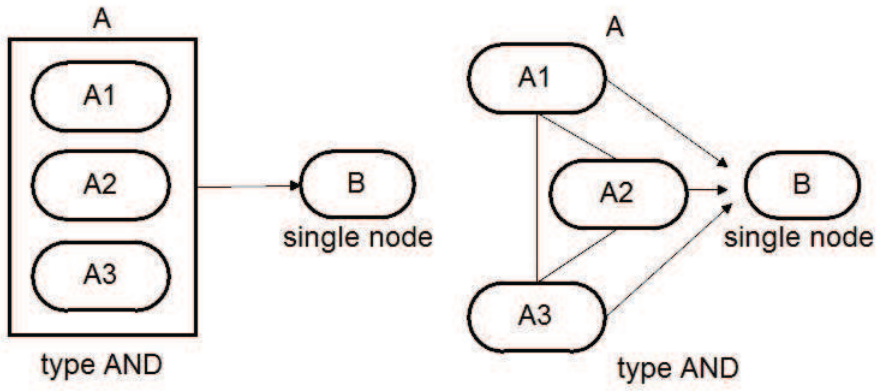

Figura 3.2: Gruppo AND[8]

- Gruppi OR (figura 3.3), ovvero gruppi che contengono elementi tra di loro alternativi, come ad esempio famiglie geniche o geni con funzioni biochimiche simili, che possono essere espansi senza dover considerare alcuna connessione con le altre proteine del gruppo;

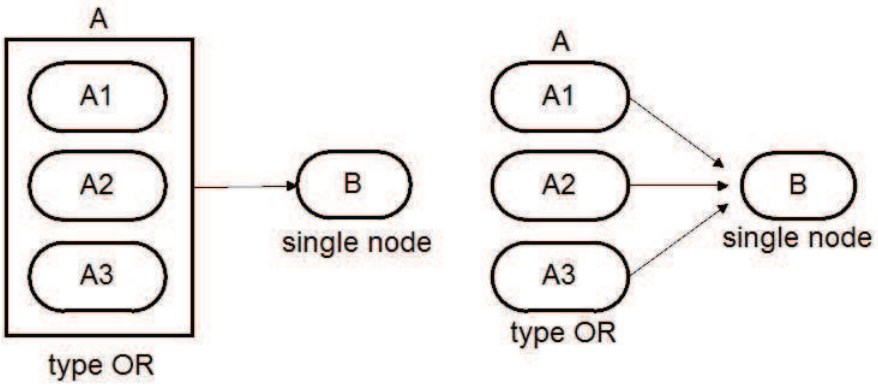

Figura 3.3: Gruppo OR[8]

I *compound* sono entità non misurabili attraverso gli *array*, la relazione tra due prodotti genici che avviene attraverso un *compound* deve essere, quindi, propagata (figura 3.4). Nel caso di più *compound* che formano una catena il segnale viene propagato fino a raggiungere i relativi prodotti genici. A un `xml file` di cui è stato riportato un esempio, corrisponde un unico *pathway*.

Nel codice formato KGML ci sono due modi di definire i nodi, composti da molti elementi:

- Gruppi AND sono definiti dal codice `type="group"`;
- Gruppi OR sono definiti dal codice `type="gene"`.

Definizione di gene:

```
<entry id="4" name="hsa:5781" type="gene"
link="http://www.kegg.jp/dbget-bin/www_bget?hsa:5781">
<graphics name="PTPN11, BPTP3, CFC, MGC14433, NS1, PTP-1D, PTP2C,
SH-PTP2, SH-PTP3, SHP2" fgcolor="#000000" bgcolor="#BFFFBF"
```

```
type="rectangle" x="496" y="235" width="46" height="17"/>
```

In questa parte di codice viene definito il gene `hsa:5781`, ovvero PTPN11 (vengono forniti anche altri nomi detti con cui di solito viene identificato il gene, ad esempio, BPTP3, CFC,...).

Definizione di gruppo OR:

```
<entry id="5" name="hsa:25 hsa:613" type="gene"
link="http://www.kegg.jp/dbget-bin/www_bget?hsa:25+hsa:613">
<graphics name="ABL1, ABL, JTK7, bcr/abl, c-ABL, p150, v-abl..."
fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle" x="301"
y="174" width="46" height="17"/>
```

Con questo codice viene definita l'entità 5 (entry id=5). Come si può notare a questa entità, o gruppo OR, appartengono due geni, il gene 25 e il gene 613.

Definizione di gruppo AND:

```
<entry id="54" name="undefined" type="group">
<graphics fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle"
x="377" y="174" width="46" height="51"/>
<component id="33"/>
<component id="34"/>
<component id="35"/>
```

Con questo tipo di codice vengono definiti i gruppi AND, ai quali non è stato dato un nome particolare, ma sono composti dalle entità specificate in `<component id="..."/>` (in questo caso il gruppo AND è formato

dalle entità 33, 34 e 35). Come si può facilmente intuire le unità che compongono il gruppo AND possono essere singoli geni ma anche gruppi OR, per questo motivo espandere i gruppi AND in un successivo momento può diventare un processo complesso.

Definizione di compound:

```
<entry id="36" name="cpd:C05981" type="compound"
link="http://www.kegg.jp/dbget-bin/www_bget?C05981">
<graphics name="C05981" fgcolor="#000000" bgcolor="#FFFFFF"
type="circle" x="506" y="174" width="8" height="8"/>
```

Qui invece viene definita l'entità 36, che in questo caso è un *compound*.

Definizione delle interazioni:

```
<relation entry1="5" entry2="54" type="PPrel">
<subtype name="activation" value="-&gt;"/>
<subtype name="phosphorylation" value="+p"/>
```

Infine con questa parte di codice vengono definite le relazioni che intercorrono tra i geni. In questo caso le entità 5 (gruppo OR), e entità 54 (gruppo AND) interagiscono tra di loro tramite una relazione *Protein-Protein*, nello specifico una reazione di fosforilazione.

Propagazione del segnale attraverso i compound:

Anche per i *compound* (figura 3.4) ci sono due modi diversi per poterli descrivere, a seconda del tipo di interazione in cui sono coinvolti:

- Interazione diretta, ovvero l'elemento A interagisce con l'elemento B attraverso il *compound* c (`type="PPrel"`, ovvero *Protein-Protein relation*);
- Interazione indiretta, ovvero l'elemento A interagisce con il *compound* c e il *compound* c interagisce con l'elemento B (`type="PCrel"`, ovvero *Protein-Compound relation*).

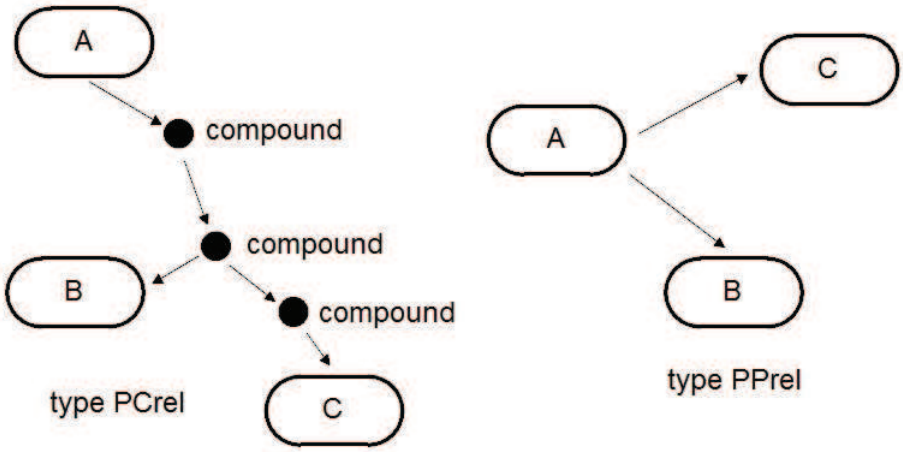

Figura 3.4 : Compound mediated signal[8]

3.1.3 Concetti base della teoria dei grafi

L'oggetto della teoria qui esposta è il grafo , ovvero una struttura composta da (N, A), dove $N = \{v_1, \ldots, vm\}$ è un insieme finito di elementi detti nodi o vertici, mentre $A \subseteq N \times N$ è un sottoinsieme di coppie ordinate di nodi dette archi. I nodi sono rappresentati con cerchi, mentre gli archi sono rappresentati con frecce che partono dal primo nodo della coppia e terminano nel secondo nodo. Si parla di grafo orientati se $v_1 \to v_2$, ovvero la coppia $(v_1, v_2) \in A$ mentre $(v_2, v_1) \notin A$, si parla, inve-

ce, di grafo non orientato se non è espressa la direzionalità dell'arco. La direzionalità espressa nel grafo orientato permette di definire le seguenti entità:

- Sono genitori di v tutti quei nodi $u \in V : (u, v) \in A$, ovvero i nodi u da cui parte un arco verso v;
- Figli di v tutti quei nodi $u \in V : (v, u) \in A$, ovvero i nodi u in cui arriva un arco da v;
- Coniugi di v tutti quei nodi u che condividono un figlio con v;
- Discendenti di v tutti quei nodi $u \in V$ per cui esiste un cammino che porta da v a u $(v \to \ldots \to u)$;
- Predecessori di v tutti quei nodi $u \in V$ per cui esiste un cammino che porta da u a v $(u \to \ldots \to v)$.

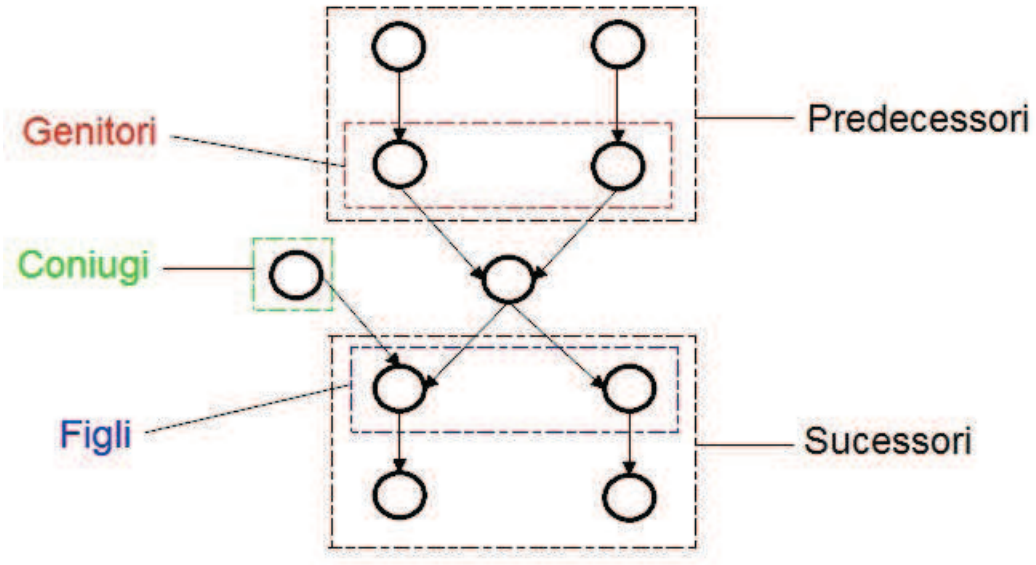

Figura 3.5:Esempio di grafo orientato aciclico

Due nodi si dicono connessi se esiste un cammino (o *path*) che li unisce, ovvero esiste una sequenza di nodi tale che ogni nodo abbia un arco che lo collega al nodo successivo. Nel caso in cui un nodo si presenti sia all'inizio che alla fine del path, questo viene definito ciclo. Se in un ciclo sono presenti degli archi orientati si parla di ciclo orientato. Si definisce, quindi, come DAG (figura 3.5) un grafo orientato aciclico (*Directed Acyclic Graph*). Dato un grafo orientato aciclico si può ottenere un grafo morale D_m aggiungendo archi non orientati tra tutte le coppie di nodi che hanno un figlio in comune e rendendo, poi, tutti gli archi presenti, non orientati. Un grafo si dice completo se A contiene tutte le coppie di elementi distinti contenuti in N. $G_b = (B, A_b)$ è un sottografo di G se $B \subseteq N$ e $A_b = A \cap (B \times B)$. Se un sottografo completo non è contenuto all'interno di nessun altro sottografo completo si chiama *clique*.

3.2 TopologyGSA

TopologyGSA è un nuovo approccio, proposto da Massa e al. [7], per l'analisi di gruppi genici definiti dai *pathway*. Questa analisi, che tiene fissa la struttura di dipendenza esistente tra i geni definita dalla topologia del *pathway*, è basata su modelli grafici Gaussiani con lo scopo di:

- Confrontare le medie e le varianze stimate del *pathway* in due condizioni sperimentali;
- Riuscire a scomporre il *pathway* in componenti più piccole che possono essere testate tra di loro al fine di trovare quelle coinvolte nel

processo di sregolazione.

3.2.1 Convertire un *network* in un modello grafico

Il primo passo che gli autori [7] eseguono per passare da un *pathway* a un modello grafico è quello di convertire il *pathway* in un grafo orientato aciclico (DAG), ovvero:

- Le reazioni di inibizione, forsforilazione e defosforilazione sono considerate come semplici archi orientati;
- Gli archi non orientati sono direzionati usando le informazioni derivanti da altri *database*;

Dopo aver convertito il *pathway* in un DAG, viene convertito in un grafo morale, secondo la metodologia esposta nel paragrafo precedente. Si assume che i dati di uno stesso *pathway* in diverse condizioni sperimentali siano realizzazioni di un modello grafico Gaussiano, ovvero:

$$M_1(G) = \left\{ Y \sim N_p(\mu_1, \Sigma_1), \Sigma_1^{-1} \in S^+(G) \right\} \tag{3.1}$$

$$M_2(G) = \left\{ Y \sim N_p(\mu_2, \Sigma_2), \Sigma_2^{-1} \in S^+(G) \right\} \tag{3.2}$$

dove p è il numero di geni, ovvero i nodi del grafo, e $S^+(G)$ è l'insieme delle matrici simmetriche definite positive con elementi nulli che corrispondono agli archi mancanti del grafo G. L'assunzione di normalità dei dati è motivata dal fatto che, come è stato dimostrato in letteratura, il logaritmo dei valori di espressione segue approssimativamente la distribuzione normale. Come è noto, le medie μ_1 e μ_2 e la matrice di variante e

covarianze, non essendo note a priori, devono essere stimate dai dati reali mediante l'utilizzo di determinati algoritmi, per approfondire si rimanda al lavoro di Lauritzen[12].

3.2.2 Verifica d'ipotesi in un modello grafico

L'espressione globale di un *pathway* può cambiare principalmente per due motivi:

1. Può cambiare la forza delle relazioni che lo definiscono;
2. Può cambiare la media dell'espressione, indipendentemente da quello che succede alla correlazione tra i geni.

Si possono definire, rispettivamente, due diversi sistemi di ipotesi sui modelli grafici, uno per testare l'ipotesi di omogeneità delle matrici di varianza e covarianza, poiché queste contengono le informazioni che riguardano le relazioni esistenti tra le entità del *pathway*, e l'altro per testare l'ipotesi di uguaglianza tra le due medie.

3.2.2.1 Test sull'uguaglianza delle matrici di concentrazione

Nell'ambito dei modelli grafici Gaussiani, per testare la forza delle relazioni esistenti tra i geni in due diverse condizioni sperimentali, si comparano le matrici di concentrazione, definite come l'inverso delle matrici di covarianza.

Il sistema d'ipotesi è del tipo:

$H_0 : \Sigma_1^{-1} = \Sigma_2^{-1}$

$H_1 : \Sigma_1^{-1} \neq \Sigma_2^{-1}$

dove Σ_1^{-1} e Σ_2^{-1} sono le matrici di concentrazione nelle due condizioni sperimentali.

Considerando $\Sigma_1^{-1} = K_1 \in S^+(G)$ e $\Sigma_2^{-1} = K_2 \in S^+(G)$ il nuovo sistema d'ipotesi diventa:

$H_0 : K_1 = K_2$

$H_1 : K_1 \neq K_2$.

Si suppone, inoltre, che le osservazioni $y_1 = (y_1^i)\, j = 1, \ldots, n_1$ provengano da una distribuzione $N_p(0,\, K_1^{-1})$ e che le osservazioni $y_2 = (y_2^j)\, j = 1, \ldots, n_2$ provengano da una distribuzione $N_p(0,\, K_2^{-1})$.

Definendo $W_i = \sum_{j=1}^{n_i} (y_i^j)(y_i^j)^T \; i = 1, 2$ la funzione di verosimiglianza è:

$$L(K_1,\, K_2) = \prod_{i=1}^{2} = (2\pi)^{-\frac{n_i p}{3}} (det\, K_i)^{\frac{n_i}{2}} e^{-\frac{1}{2} tr(K_i W_i)} \tag{3.3}$$

$\hat{K}_1$e $\hat{K}_2$ possono essere stimate tramite cacolo diretto se il grafo è scomponibile oppure tramite l'algoritmo proposto da Lauritzen[12]. Sotto H_0, l'algoritmo produce una stima della matrice di covarianza comune $(\hat{\Sigma})$ a partire dalla matrice di covarianza *pooled*

$$S = (n_1 + n_2 - 2)^{-1} \{(n_1 - 1)\, S_1 + (n_2 - 1)\, S_2\} \tag{3.4}$$

Mentre sotto l'ipotesi alternativa $\hat{\Sigma}_1$ e $\hat{\Sigma}_2$ sono stimate usando le matrici di covarianza campionaria

$$S_1 = (n_1 - 1)^{-1} W_1 \tag{3.5}$$

$$S_2 = (n_2 - 1)^{-1} W_2 \tag{3.6}$$

Una volta ottenute le stime elencate sopra si calcola il test rapporto di verosimiglianza:

$$\Lambda = \frac{L_{H_0}(\hat{K}_1, \hat{K}_2)}{L_{H_1}(\hat{K}_1, \hat{K}_2)} = \frac{L_{H_0}(\hat{K})}{L_{H_1}(\hat{K}_1, \hat{K}_2)} \tag{3.7}$$

Dopo opportune semplificazioni [12]si arriva a definire:

$$-2log\Lambda = \sum_{i=1}^{2} n_i log\left(\frac{det\hat{K}_i}{det\hat{K}}\right) \tag{3.8}$$

che si distribuisce asintoticamente come χ^2_{r+p}, dove r è il numero di archi di G. Se il test rifiuta l'ipotesi nulla può essere interessante andare a trovare l'origine delle differenze nelle matrici di concentrazione. Se il grafo è scomponibile è possibile andare a ripetere l'analisi in ognuna delle *cliques*.

3.2.2.2 Test per l'uguaglianza delle medie

Per testare la differenziale espressione del *pathway* in due diverse condizioni sperimentali si considera il seguente sistema d'ipotesi:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Naturalmente il test usato per testare questa ipotesi dipenderà dall'esito del test precedente. Se l'ipotesi nulla di uguaglianza delle matrici di covarianza non è rifiutata , ovvero $\Sigma_1 = \Sigma_2$, si userà l'approccio MANOVA,

ovvero una versione multivariata generalizzata dell'analisi della varianza univariata (ANOVA) [13]. Nel caso in cui l'ipotesi nulla venga rifiutata si usa l'usuale test per l'uguaglianza delle medie in due condizioni sperimentali con diverse matrici di concentrazione.

3.2.2.3 Decomposizione del grafo in cliques

L'utilizzo dei modelli grafici, nel contesto della *gene set analysis*, ci consente, dopo aver moralizzato il pathway (vedi paragrafo 3.1.3), di scomporre tale grafo in sottografi completi, detti *cliques*. Questa caratteristica risulta utile se i test sulle medie e sulle matrici di concentrazione (3.29) rifiutano l'ipotesi nulla in quanto ci permette di capire l'origine di queste differenze. Il test può essere dunque ripetuto per ognuna delle *cliques*, identificate all'interno del grafo, indipendentemente dalle altre. In questo modo, la stima di massima verosimiglianza delle matrici di concentrazione relative alle variabili che compongono una stessa *clique* è ottenuta con i dati disponibili relativi alle *cliques* e quindi non c'è necessita di alcuna marginalizzazione.

3.3 SPIA (*Signaling pathway Impact Analysis*)

Un nuovo approccio chiamato SPIA, è stato proposto da Tarca et al.[15]. Questo approccio combina due tipi di evidenze:

- La sovra-rappresentazione dei geni differenzialmente espressi in un dato *pathway*;
- Le anomale perturbazioni del *pathway*, misurate come cambiamenti nei livelli di espressione nella topologia del *pathway*.

Per catturare questi due aspetti vengono definiti due valori di probabilità indipendenti, P_{NDE} e P_{PERT} (un esempio pratico è mostrato in figura 3.6).

$$P_{NDE} = P(X \geq N_{de} \mid H_0) \tag{3.9}$$

misura la probabilità di ottenere un numero di geni differenzialmente espressi maggiori di quelli osservati in un dato *pathway*. H_0 è l'ipotesi nulla secondo cui i geni differenzialmente espressi che appaiono nel *pathway* sono dovuti solamente al caso. I valori di P_{NDE} sono ottenuti supponendo che NDE, ovvero il numero di geni DE nel *pathway* analizzato, segua una distribuzione ipergeometrica di parametri m, n, k, dove

- m è il numero di tutti i geni presenti nell'array;
- n è il numero di geni all'interno dell'*array* che non appartengono al *pathway*;
- k il numero totale di geni differenzialmente espressi.

Per il calcolo della seconda probabilità P_{PERT}, che si basa su una misura della perturbazione in ciascun *pathway*, è necessario definire un fattore genico di perturbazione $PF(g_i)$.

$$PF(g_i) = \triangle E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)} \tag{3.10}$$

Il primo termine dell'equazione rappresenta una misura normalizzata dei cambiamenti nei valori di espressione del gene g_i, *log-fold change* se si stanno confrontando due condizioni sperimentali. Il secondo termine è la somma dei fattori di perturbazione dei geni g_j, predecessori di g_i, normalizzati per il numero di geni discendenti di ciascuno di questi geni indicati con $N_{ds}(g_j)$. Il parametro β_{ij} quantifica la forza delle interazioni presenti tra i geni g_j e g_i, con $\beta = 0$ nel caso in cui l'arco non esista. L'equazione (3.10) descrive il fattore genico di perturbazione per il gene g_i come una funzione lineare dei fattori di perturbazione relativi ai geni che compongono tutto il *pathway*. Una volta calcolato il fattore genico di perturbazione si calcola l'accumulo di perturbazione netta (*net perturbation accumulation*), per ogni singolo gene, come la differenza tra PF e il suo *log fold-change* osservato

$$Acc(g_i) = PF(g_i) - \triangle E(g_i) \tag{3.11}$$

Questo fattore serve per assicurarsi che i geni differenzialmente espressi, che non sono legati a nessun altro gene, non contribuiscano alla seconda evidenza, dal momento che sono già stati considerati nell'analisi di arricchimento catturata dal primo termine dell'equazione (3.10). La probabilità di osservare un valore di perturbazione accumulata totale del

pathway (T_A) maggiore di t_A è

$$P_{PERT} = P(T_A \geq t_A \mid H_0) \tag{3.12}$$

dove $t_A = \sum_i Acc(g_i)$. Per determinare in modo empirico la distribuzione nulla di T_A viene utilizzato un approccio di ricampionamento di tipo *Bootstrap*[15]. Una probabilità di sintesi, che combina le due probabilità P_{NDE} e P_{PERT}, è data dalla probabilità globale P_G che è utilizzata per ordinare i *pathway*s e per testare l'ipotesi che il *pathway* sia realmente perturbato nelle condizioni di studio considerate. Se definiamo con c_i il prodotto delle due probabilità indipendenti P_{NDE} e P_{PERT}, ovvero

$$c_i = P_{NDE}(i) \cdot P_{PERT}(i) \tag{3.13}$$

nel caso in cui H_0 sia vera, si può dimostrare che la probabilità di osservare dei valori di p-value per P_{NDE} e P_{PERT} talmente bassi da far risultare il loro prodotto più piccolo del valore realmente osservato è riassunta dal valore di P_G definito come

$$P_G = c_i - c_i \cdot ln(c_i) \tag{3.14}$$

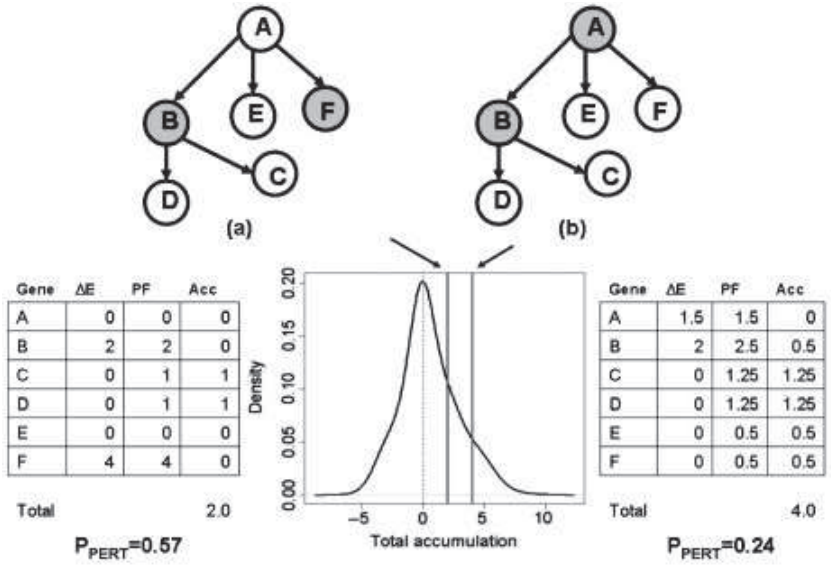

Figura 3.6: Esempio di analisi SPIA [15]

Va sottolineato il fatto che entrambe le probabilità P_{NDE} e P_{PERT} sono indipendenti dalla grandezza del *pathway* per la natura con cui sono state costruite. P_{NDE} misura la probabilità di osservare un numero di geni differenzialmente espressi, dovuti solamente al caso, maggiore del numero osservato e il numero di geni DE dovuti al caso aumentano all'aumentare della dimensione del *pathway*. P_{PERT} è ottenuta con un approccio bootstrap calcolato su ogni singolo *pathway* e risulta quindi indipendente dalla sua dimensione. P_G, essendo combinazione di due probabilità, può essere usato non solo per ordinare i *pathway*s, ma anche per scegliere un livello desiderato per l'errore di primo tipo. SPIA è un valido approccio anche sotto il profilo della specificità. E' stato dimostrato in letteratura [15], tramite studi di simulazione, che SPIA non produce, quando H_0 è vera, un numero di falsi positivi superiore a quelli definiti dalla soglia α. Nelle analisi su dati reali però, per confrontare diversi metodi di analisi

su uno stesso insieme di *pathway*s, non vengono utilizzate le misure di sensibilità, specificità e le curve ROC in quanto non si riesce a stabilire con esattezza il livello di coinvolgimento di un dato *pathway* in una determinata condizione sperimentale. Per questo motivo, quando si parla di confronti tra metodi diversi di analisi ci si riferisce principalmente a confronti in termini di numero di *pathway*s risultati significativi e quanto bene questi *pathway*s di adattano alle conoscenze biologiche esistenti.

4 Dati reali e risultati

4.1 I dati

In letteratura [11] è ormai noto quali sono i meccanismi genici che portano alla trasformazione maligna delle cellule coinvolte nella leucemia linfoblastica acuta (ALL). Nel 25-30% dei pazienti adulti affetti da questa malattia si è notato come molti geni fossero associati a differenziazioni nelle cellule T, nelle quali i geni recettori mostravano la presenza di riarrangiamenti clonali, mentre nel rimanente 70-75% di pazienti si notavano differenziazioni a livello delle cellule B. In particolare, nelle cellule B, sono frequenti le traslocazioni cromosomiche e i riarrangiamenti molecolari, con un'incidenza maggiore negli adulti piuttosto che nei bambini. La leucemia mieloide [9] cronica è una delle prime malattie per cui sia stato possibile individuare una specifica anormalità cromosomica quale causa della malattia. La traslocazione del gene Abelson (ABL) dal cromosoma 9 a una regione del cromosoma 22, denominata *breakpoint cluster region* (BCR), con la formazione di un gene chimera BCR/ABL, è un difetto cromosomico presente nel 95% di tutti i pazienti affetti da leucemia mieloide cronica e nel 30% circa di pazienti adulti affetti da leu-

cemia linfoblastica acuta (ALL). I dati utilizzati in questo elaborato sono i dati B-cell[1]. Essi provengono da tecnologia *Affymetrix* a singolo canale (*Affymetrix* U95Av2, Santa Clara CA). I campioni leucemici ($n = 128$) sono stati selezionati da un totale di 431 pazienti, di età compresa tra i 15 e i 58 anni, coinvolti in uno studio *follow-up* dall'università "La Sapienza" di Roma tra il 1996 e il 2000. Tra questi 128 pazienti, 95 presentano differenziazioni nelle cellule B e 33 nelle cellule T. Il campione finale di pazienti su cui verranno svolte le analisi comprenderà 37 campioni che sono risultati positivi, nell'analisi citogenetica, ai riarrangiamenti molecolari BCR/ABL, e 41 campioni, detti NEG, che non manifestano riarrangiamenti molecolari evidenti. Le analisi condotte su questi dati sono state sviluppate tramite il *software* statistico R[2]. Come prima cosa, poiché i dati provengono da tecnologia *Affymetrix*, e sono quindi presentati come file .CEL, sono stati caricati utilizzando il comando `ReadAffy()` della libreria `Affy`[3], che consente di leggere questo tipo di file e di posizionare i dati in un *AffyBatch*, nel quale sono presenti le principali informazioni di sintesi dei dati. Nello specifico sono presenti 78 campioni per un totale di 12625 geni. Il passo successivo è dato dall'annotazione dei dati. L'annotazione è il processo attraverso il quale ad ogni *probeset* viene l'identificativo di ogni gene. Questa procedura utilizza i *custom definition file* (CDF), file di mappatura biunivoci *probeset-gene*. Nello specifico si è usata il CDF della piattaforma *Affymetrix* utilizzata per gli esperimenti:

[1] Dal sito di Bioconductor: http://www.bioconductor.org/help/publications/2003/Chiaretti/chiaretti2/

[2] Scaricabile liberamente dal sito http://www.r-project.org/

[3] Dal sito di Bioconductor: http://www.bioconductor.org/packages/2.9/bloc/html/affy.html

U95Av2 [4]. A questo punto sì è potuto normalizzare i dati attraverso una normalizzazione *Robust Multi-array Analysis (RMA)* e quantile [3, 4]. Il modello RM, per la normalizzazione dei dati, prevede i seguenti passi:

- Correzione del segnale per il *background*, con lo scopo di aggiustare le intensità dei PM al fine di rimuovere l'effetto dovuto al *background*;
- Normalizzazione quantile-quantile, che ha lo scopo di rendere uguali le distribuzioni dell'intensità dei *probe* per ciascun *array*;
- Utilizzo di un modello robusto ad effetti fissi (*probeset*) tra *array* per stimare il valore di espressione.

4.2 Il *pathway Chronic Myeloid Leukemia*

La leucemia mieloidale cronica è originata in una cellula ematopoietica pluripotente nel midollo osseo. La cellula pluripotente è in grado di dare origine a più popolazioni cellulari, in questo caso del midollo osseo. La principale caratteristica di questa malattia è data dal numero sempre crescente di granulociti nel sangue. In questo elaborato è stato utilizzato il *pathway Chronic Myeloid Leukemia*, presentato in figura 3.1[5]. Per avere una visione più chiara delle vie di segnale che compongono un *pathway* e per avvicinarsi maggiormente alla sua raffigurazione originale la figura 3.1 in alcuni casi presenta alcuni nodi che sono duplicati. Il grafo originale

[4] Dal sito: http://genecards.weizmann.ac.il/geneannot/customcdf.shtml
[5] Dal sito: http://www.genome.jp/kegg-bin/show_pathway?hsa05220

e semplificato su cui sono state svolte le analisi è riportato in Appendice (Figure 1, 3).

Il *pathway* visualizzato offre molte informazioni oltre a quelle puramente grafiche già introdotte nel paragrafo 3.1. Se si clicca sul rettangolo in alto a sinistra contenente il nome del *pathway* viene restituita una pagina *web* contenente:

- La descrizione del *pathway*;
- Il nome del *pathway*;
- Una descrizione riassuntiva del processo biologico mostrato nella mappa;
- Un link a *KEGG DISEASE* in cui vengono date informazioni riguardanti la malattia a cui i geni sono associati;
- La lista dei nodi coinvolti nel *pathway*, sia geni che proteine, rappresentate graficamente con dei rettangoli;
- La lista dei compound presenti nel *pathway*, rappresentati graficamente con dei cerchi;
- La lista delle referenze utilizzate per la creazione del disegno del *pathway*;
- La lista dei *pathways* collegati con quello in esame.

Inoltre, tornando alla mappa del *pathway*, se ci si posiziona con il mouse sopra una delle entità, viene visualizzato il nome del gene e l'identificarivo

ID dello stesso[6]. Si nota, inoltre, che in alcuni casi vengono restituiti più ID genici, questo sta a segnalare che si è in presenza di gruppi AND o gruppi OR, ma per capire a quale delle due tipologie di gruppi ci si riferisce è necessario fare riferimento al codice KGML.

4.2.1 Il *pathway* semplificato

Riferendosi al pathway introdotto nel paragrafo precedente si è analizzato il KGML e si sono individuati i gruppi AND, OR e le loro eventuali nidificazioni (vedi tabella 4.1 e figura 4.2). L'idea alla base di questo elaborato è quella di semplificare il *pathway* dal punto di vista topologico e di valutare quale sia la più appropiata sintesi statistica dei dati. Questa sintesi è stata pensata, oltre che per semplificare la struttura del *pathway*, anche per rendere computazionalmente più veloce l'analisi. Come sintesi sono state proposte:

- La media aritmetica come livello di espressione di sintesi dei gruppi OR e dei gruppi AND ("Medie" (ME));
- Il gene che presenta massima differenziale espressione in media tra le due condizioni sperimentali come sintesi dei gruppi OR, mentre per i gruppi AND la media aritmetica ("Media e Massima Differenziale Espressione" (MMDE));

[6]Per maggiori informazioni relative a ogni singolo gene vedere http://www.ncbi.nlm.nih.gov/gene/

- Il gene che ha massima differenziale espressione in media tra le due condizioni sperimentali sia per i gruppi OR che per i gruppi AND ("Massima Differenziale Espressione" (MDE));
- La prima componente principale, sia per i gruppi OR che per i gruppi AND ("Componente Principale" (CP)).

In questo modo si passa da una situazione rappresentata in figura 4.1 in cui sono espansi tutti i gruppi OR e tutti i gruppi AND, a una situazione semplificata rappresentata in figura 4.2 in cui ad un gruppo corrisponde un nodo.

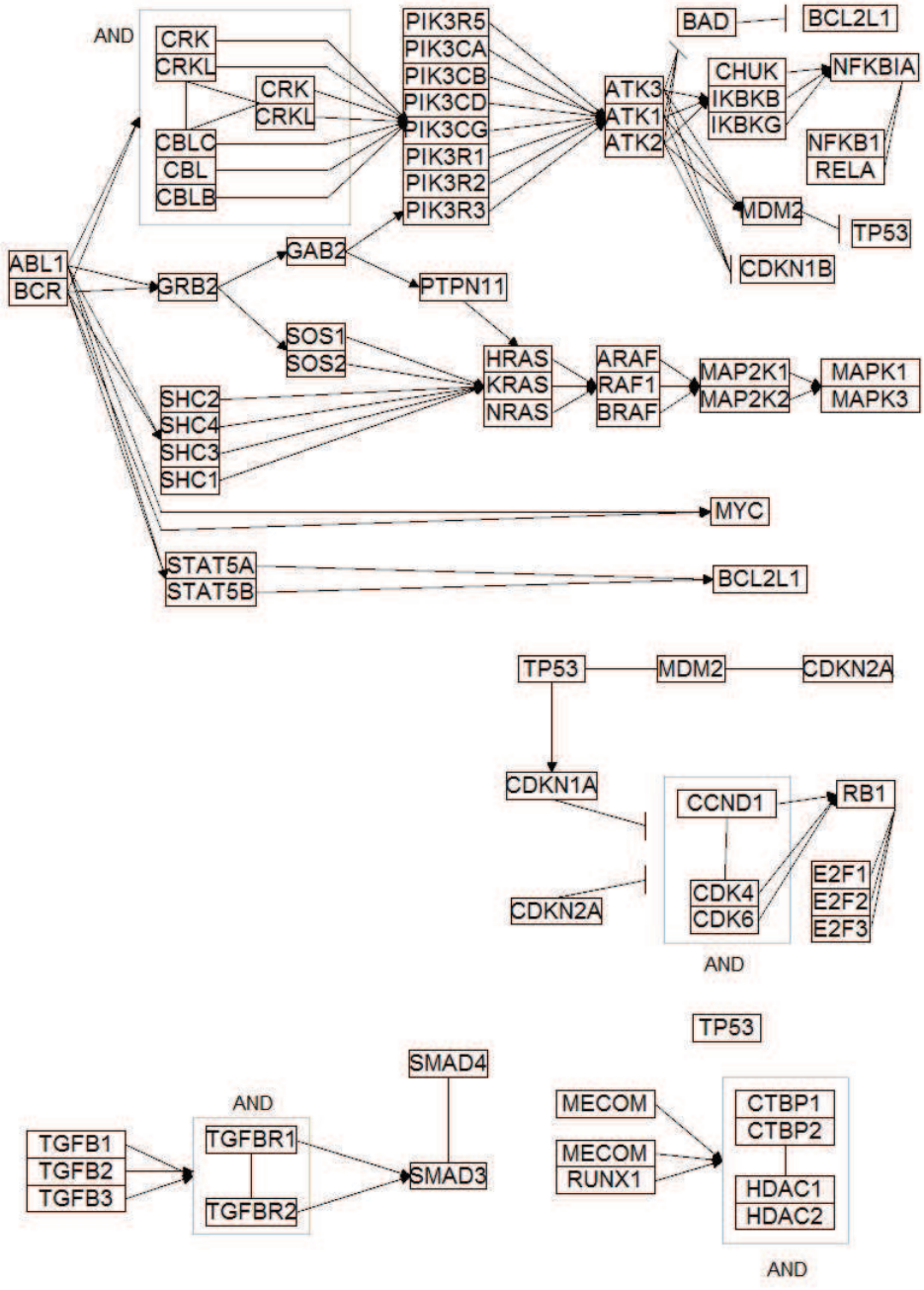

Figura 4.1: Il pathway espanso

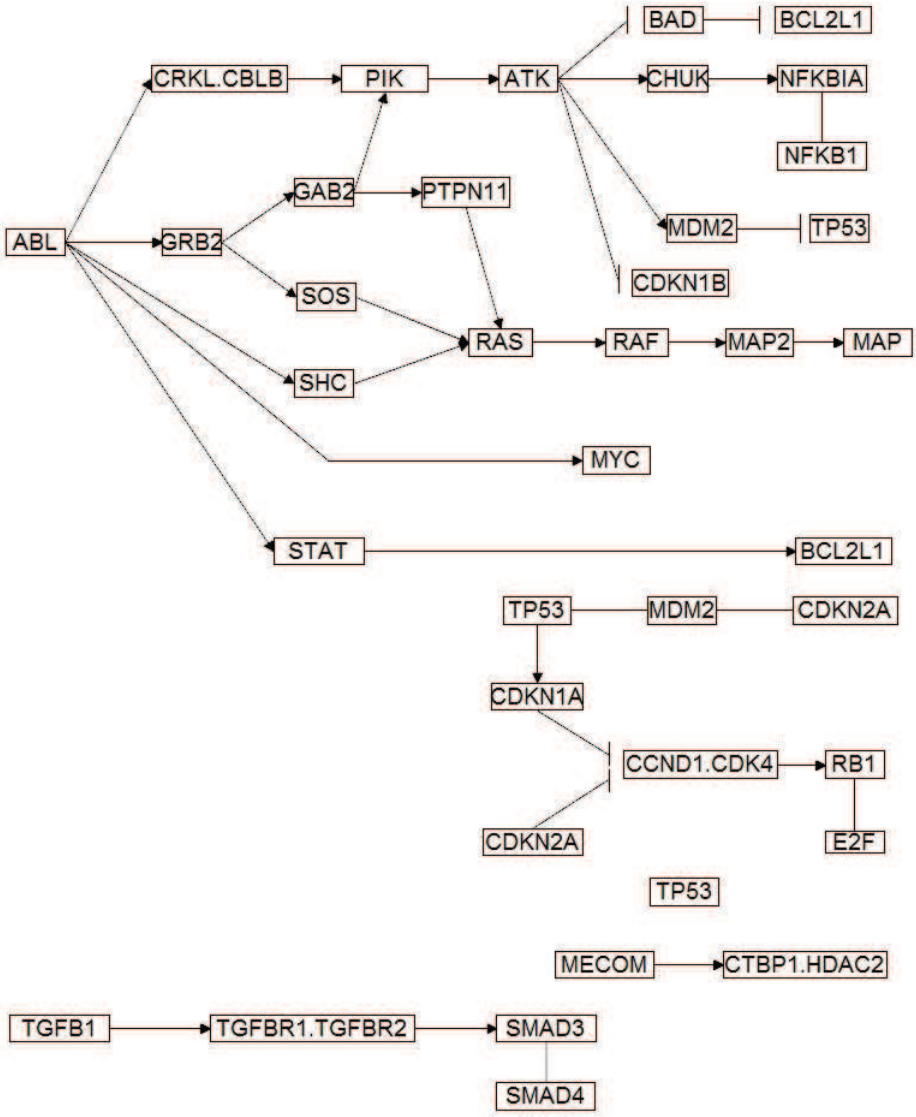

Figura 4.2: Il pathway semplificato

4.2.2 I quattro dataset semplificati

I risultati delle quattro situazioni rappresentate nel paragrafo precedente sono state tutte confrontate con i risultati ottenuti in condizioni normali, ovvero la situazione in cui il *pathway* è completamente espanso. La matrice dei dati originale, ottenuta dai dati B-cell[7] opportunamente annotati e normalizzati, è una matrice di 8384 righe (geni) e 78 colonne, di cui 37 riferite ai campioni positivi alla traslocazione BCR, nel seguito BCR, e 41 riferite ai campioni negativi , nel seguito NEG. Il dataset "Iniziale" (da qui in poi chiamata "MI"), è stato ottenuto selezionando dalla matrice dei dati iniziale, contenente 8384 geni (nodi), gli elementi del *pathway Chronic myeloid leukemia* utilizzando il pacchetto `graphite` [8] di R. In questo modo "MI" diventa una matrice 68x78, dove nel senso delle righe si trovano i nomi dei nodi del *pathway*, mentre nel senso delle colonne troviamo i campioni nelle due situazioni sperimentali. Partendo da questa matrice sono state create le quattro matrici introdotte nel paragrafo precedente, che avranno dimensione ridotta rispetto a "MI". Nello specifico saranno tutte matrici 34x78, con un numero di righe, ovvero di nodi, pari esattamente alla metà rispetto a "MI".

"ME", è una matrice in cui sono stati presi i geni di ogni singolo gruppo OR, anche di quelli a loro volta contenuti nei gruppi AND, ed è stata fatta la media aritmetica. Stessa operazione è stata successivamente svolta per i gruppi AND.

"MMDE" è una matrice in cui i geni che compongono i gruppi OR sono

[7] http://www.bioconductor.org/help/publications/2003/Chiaretti/chiaretti2/
[8] http://www.bioconductor.org/packages/release/bioc/html/graphite.html

rappresentati dal gene che presenta una maggiore differenziale espressione in media tra il gruppo BCR e il gruppo NEG, mentre per quanto riguarda i gruppi AND si è deciso operare allo stesso modo della matrice precedente.

"MDE" è una matrice in cui, inizialmente si prendono come rappresentanti dei gruppi OR i geni che presentano una maggiore differenziale espressione in media tra il gruppo BCR e il gruppo NEG, e successivamente si è svolta la stessa operazione per i gruppi AND.

"CP" è la matrice in cui i livelli di espressione dei gruppi OR e dei gruppi AND vengono rappresentati dalla prima componente principale.

Nella tabella 4.1 vengono riportati i nomi assegnati ai gruppi OR e AND, e i geni che li compongono.

	Nomi gruppi	Geni che li compongono
OR	ABL	ABL1, BCR
	CRKL	CRK, CRKL
	CBLB	CBL, CBLB
	PIK	PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2, PIK3R3
	ATK	ATK3, ATK1, ATK2
	CHUK	CHUK, IKBKB, IKBKG
	NFKB1	NFKB1, RELA
	SOS	SOS1, SOS2
	RAS	HRAS, KRAS, NRAS
	RAF	ARAF, RAF1, BRAF
	MAP2	MAP2K1, MAP2K2
	MAP	MAPK1, MAPK3
	SHC	SHC2, SHC3, SHC1
	STAT	STAT5A, STAT5B
	CDK4	CDK4, CDK6
	E2F	E2F1, E2F2, E2F3
	MECOM	MECOM, RUNX1
	CTBP1	CTBP1, CTBP2
	HDAC2	HDAC1, HDAC2
	TGFB	TGFB1, TGFB2, TGFB3
AND	CRKL.CBLB	CRKL, CBLB
	TGFBR1.TGFBR2	TGFBR1, TGFBR2
	CCND1.CDK4	CCND1, CDK4
	CTBP1.HDAC2	CTBP1, HDAC2

Tabella 4.1: **Nomi assegnati ai gruppo OR e AND e i geni che li compongono**

4.3 I geni differenzialmente espressi

La prima cosa che è stata fatta nelle cinque matrici è stata l'identificazione dei geni differenzialmente espressi. Come primo passo è stata creata la matrice del disegno,

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \dots & \dots \\ 1 & 1 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{bmatrix} \quad 37 - esima\ posizione \tag{4.1}$$

Poi con l'utilizzo del pacchetto `Limma`[9], è stato stimato un modello lineare per ogni singolo gene, in ognuna delle matrici considerate, attraverso la funzione `lmFit()`, su cui poi sono state stimate la statistica t-moderata (2.19) e i log-odds della differenziale espressione utilizzano la funzione `ebayes()`. Una volta ottenute queste stime, si sono visualizzati i risultati utilizzando la funzione `topTable()`. Qui di seguito vengono riportate delle tabelle contenenti le principali statistiche di sintesi ottenute per i geni differenzialmente espressi individuati in ognuno delle matrici.

[9] E' possibile scaricare il pacchetto all'indirizzo web: http://bioconductor.org/packages/release/bioc/html/limma.html

ID	logFC	AveExpr	t	p.value
ABL1	-0.7560	7.2740	-7.8552	$1.6160e^{-11}$
NFKBIA	-0.7120	10.5990	-3.4027	0.00105
NFKB1	-0.3640	8.6263	-2.8904	0.00496
CDK4	0.2775	7.0711	2.6508	0.00969
BCR	0.2644	6.7020	2.6042	0.01099
MDM2	-0.1503	3.5198	-2.6011	0.01108
RB1	0.1976	5.1338	2.4636	0.01592
MECOM	-0.1276	3.9458	-2.1640	0.03347
E2F1	0.1158	5.9601	2.1565	0.03407
STAT5A	-0.2235	7.0909	-1.9467	0.05511
MYC	0.1620	7.1261	1.9066	0.06019
MAP2K2	0.1559	7.6364	1.7961	0.07628
CBLB	-0.2378	6.6899	-1.7104	0.09109
BAD	0.2074	5.3367	1.6962	0.09376

Tabella 4.2: **Geni differenzialmente espressi in "MI"**

ID	logFC	AveExpr	t	p.value
NFKBIA	-0.7120	10.5990	-3.4048	0.00104
ABL	-0.2458	6.9880	-3.0979	0.00269
MDM2	-0.1503	3.5198	-2.6188	0.01056
RB1	0.1976	5.1338	2.4724	0.01556
NFKB1	-0.1927	7.9307	-2.1294	0.03632
MYC	0.1620	7.1261	1.9127	0.05939
BAD	0.2074	5.3367	1.6990	0.09324

Tabella 4.3: **Geni differenzialmente espressi in "ME"**

ID	logFC	AveExpr	t	p.value
ABL	-0.7560	7.2740	-7.8339	$1.6861e^{-11}$
NFKBIA	-0.7120	10.5990	-3.4134	0.00101
NFKB1	-0.3640	8.6263	-2.8915	0.00493
CCND1.CDK4	0.1590	6.1973	2.5688	0.01206
MDM2	-0.1503	3.5198	-2.5607	0.01232
RB1	0.1976	5.1338	2.4490	0.01651
MECOM	-0.1276	3.9458	-2.1321	0.03606
STAT	-0.2235	7.0909	-1.9456	0.05520
MYC	0.1620	7.1261	1.8974	0.06137
CRKL.CBLB	-0.1421	6.1440	-1.8801	0.06372
MAP2	0.1559	7.6364	1.7752	0.07965
BAD	0.2074	5.3367	1.6964	0.09369

Tabella 4.4: **Geni differenzialmente espressi in "MMDE"**

ID	logFC	AveExpr	t	p.value
ABL	-0.7560	7.2740	-7.8158	$1.8283e^{-11}$
NFKBIA	-0.7120	10.5990	-3.4118	0.00101
NFKB1	-0.3640	8.6263	-2.8876	0.00499
CCND1.CDK4	0.1590	6.1973	2.6415	0.00992
MDM2	-0.1503	3.5198	-2.5447	0.01285
RB1	0.1976	5.1338	2.4409	0.01685
MECOM	-0.1276	3.9458	-2.1193	0.03716
STAT	-0.2235	7.0909	-1.9425	0.05559
MYC	0.1620	7.1261	1.8918	0.06212
MAP2	0.1559	7.6364	1.7662	0.08116
CRKL.CBLB	-0.2378	6.6899	-1.7105	0.09104
BAD	0.2074	5.3367	1.6640	0.09415

Tabella 4.5: **Geni differenzialmente espressi in "MDE"**

ID	logFC	t	p.value
NFKBIA	-0.7120	-3.3666	0.00118
ABL	-0.7560	-2.4576	0.01619
MDM2	-0.1503	-2.4408	0.01690
RB1	0.1976	2.3747	0.02001
MECOM	-0.1276	1.8954	0.06172
MYC	0.1620	1.8438	0.06899
NFKB1	-0.3640	-1.8046	0.07498

Tabella 4.6: **Geni differenzialmente espressi in "CP"**

Come si può notare nella tabella 4.2, nella matrice "MI" vengono individuati, a un livello $\alpha = 0,1$ per il *p-value* non corretto, soglia che verrà mantenuta anche per tutte le altre situazioni, 14 geni differenzialmente espressi. Relativamente a "ME" vengono individuati solo 7 geni differenzialmente espressi; va notato che il gruppo OR "ABL", risultato della media dei geni ABL1 e BCR, nasconde parzialmente l'effetto del primo che, come si vede in tabella 4.3, risulta essere altamente significativo. Usando la media come sintesi per i valori di espressione dei due gruppi si nota come non risulti significativo alcun gruppo AND e solo 2 tra i gruppi OR (ABL e NFKB1). Nella tabella 4.2 risultano significativi geni che formano gruppi OR come CDK4 (significativo a un livello $\alpha = 0,01$), MECOM, E2F1 (significativi a livello $\alpha = 0,05$), STAT5A e CBLB (significativi a un livello $\alpha = 0,1$), che non sono, però, stati identificati utilizzando la media come sintesi. Nella Tabella 4.4 sono riportati i geni differenzialmente espressi trovati quando viene considerato come rappresentativo, per il gruppo OR, il gene che presenta massima differenziale espressione, mentre per i gruppi AND viene utilizzata come sintesi sempre la media aritmetica. In questo caso si può notare come il

numero di geni che viene identificato sia superiore rispetto alla situazione precedente. Sono identificati come differenzialmente espressi 12 geni, di cui 5 gruppi OR (ABL, NFKB1, MECOM, STAT e MAP2) e 2 gruppi AND (CCND1.CDK4 e CRKL.CBLB). In particolare si nota che il gruppo ABL è tornato a essere altamente significativo, e inoltre si vede come i geni che sono stati identificati corrispondano a quelli identificati in "MI", con la sola eccezione del gene E2F1 (in questo caso appartenente al gruppo OR denominato E2F) che non viene identificato. Un'altra differenza sta nel fatto che il singolo gene CBLB, se considerato da solo, risulta avere una significatività minore ($p.value$=0.09110) di quella che invece si trova considerandolo all'interno del gruppo AND "CRKL.CBLB" ($p.value$=0.06372). Nella tabella 4.5 sono riportati i geni differenzialmente espressi trovati quando viene considerato come rappresentativo, per entrambe le tipologie di gruppi, il gene che presenta massima differenziale espressione in media. I risultati ottenuti sono molto simili a quelli trovati nel caso precedente. Vengono identificati, come differenzialmente espressi, gli stessi 12 geni, tra i quali 5 gruppi OR e 2 gruppi AND. In questo caso si nota come il gruppo AND "CRKL.CBLB" torni ad avere una significatività molto simile a quella mostrata dal gene CBLB nel caso in cui venga considerato da solo. Nella tabella 4.6, infine, sono riportati i geni identificati come differenzialmente espressi quando si considera come sintesi dei valori di espressione la prima componente principale. E' evidente come questa procedura nasconda la differenziale espressione dei gruppi AND e OR. In totale vengono identificati 7 geni, di cui 3 gruppi OR e nessun gruppo AND. In più si nota anche che la significatività vie-

ne ridotta molto in tutti i geni identificati, soprattutto se si considera il gruppo ABL che passa da un livello di significatività prossimo a 0 a un p.value=0.016191.

4.4 L'analisi topologica

In questo paragrafo vengono presentati i risultati dell'analisi topologica, svolta utilizzando *TopologyGSA* all'interno del pacchetto `graphite`[10]. In questa analisi vengono utilizzati i modelli grafici per testare le componenti del *pathway*, focalizzando l'attenzione sulle componenti del pathway moralizzato e triangolarizzato (vedi figura 4 in Appendice) coinvolte nel processo di deregolazione. Per tutte le analisi presentate in questo paragrafo è stato utilizzato come livello di significatività dei test un livello α pari a 0.05. Nella prima analisi svolta, i cui risultati sono rappresentati nella tabella 4.7, è stata testata l'ipotesi di uguaglianza delle matrici di concentrazione, nelle due condizioni sperimentali, nel *pathway* espanso. Si può notare come il valore osservato del livello di significatività della statistica test (`alpha.obs`) risulti altamente significativo. Per quanto riguarda il *pathway* semplificato, utilizzando una delle metodologie esposte precedentemente, i risultati più significativi sono dati dalle metodologie "MMDE" e "MDE". Nel caso in cui si utilizzi come sintesi per i gruppi AND e OR la componente principale l'ipotesi di uguaglianza delle matrici di concentrazione viene rifiutata a un livello soglia di 0.05 (`alpha.obs = 0.1970`).

[10] http://www.bioconductor.org/packages/release/bioc/html/graphite.html

	MI	ME	MMDE	MDE	CP
alpha.obs	$1.9892e^{-07}$	0.0148	0.0004	0.0003	0.1970
check	TRUE	TRUE	TRUE	TRUE	FALSE
lambda.obs	374.2507	108.1133	126.7783	127.087	91.1798
lambda.theo	279.2876	96.2167	96.2167	96.2167	96.2167

Tabella 4.7: **Analisi sulle varianze**

La seconda analisi, i cui risultati sono riportati nella tabella 4.8, mira a testare l'uguaglianza dei livelli di espressione nelle due condizioni sperimentali (BCR e NEG), condizionatamente ai risultati ottenuti con l'analisi sulle matrici di concentrazione. L'ipotesi di uguaglianza dei livelli di espressione, a un livello di significatività pari a 0.05, viene rifiutata quando si considera la matrice "MI". Questa viene rifiutata anche quando si considera una qualsiasi sintesi esposta sopra, sebbene è evidente che quando si considera come sintesi per i gruppi OR la massima differenziale espressione il valore del p.value osservato sia nullo.

	MI	ME	MMDE	MDE	CP
alpha.obs	0.0001	0.0078	0	0	0.0131
Number of Nodes	65	34	34	34	34
Number of Edges	178	41	41	41	41

Tabella 4.8: **Analisi sulle medie**

4.4.1 Le analisi sulle *cliques*

In questo paragrafo vengono esposti i risultati derivanti dall'analisi svolta sulle singole *cliques*, individuate dopo aver trasformato il grafo del *pathway* in un DAG, e dopo averlo moralizzato, e triangolarizzato se necessario, così da rendere possibile la sua decomposizione. Per condurre

questa analisi si è utilizzato il pacchetto `topologyGSA`[11] e in particolare la funzione `clique.var.test()` che permette di confrontare le matrici di concentrazione, nelle due condizioni sperimentali, di tutte le *cliques* del *pathway*. Successivamente, utilizzando la funzione `clique.mean.test()` è stata testata l'uguaglianza delle medie, nelle due condizioni sperimentali, per tutte *cliques* del *pathway*, condizionatamente ai risultati ottenuti dal test condotto sulle *cliques* precedentemente. In questo paragrafo vengono confrontate le *cliques* significative tra i vari metodi di semplificazione e il *pathway* espanso. Il *pathway* originale però avendo un numero maggiore di nodi e di *edges* non avrà lo stesso numero e la stessa composizione delle *cliques* del *pathway* semplificato. Per questo motivo ci limiteremo a confrontare i metodi sulla base dei geni contenuti nelle *cliques* significative. Per fare un esempio se la *clique* X nel *pathway* originale è composta dai geni 1, 2 e 3 e la *clique* Y del *pathway* semplificato è composta dal gruppo 1-2 e il gene 3, queste sono considerate uguali. Nella tabella 4.9 sono riportate le *cliques*, significative a un livello $\alpha = 0.05$, della matrice "MI" e i relativi valori di `alpha obs.` per i due test. Ovvero `alpha obs.1` descrive il livello di significatività osservato quando viene testata l'ipotesi di uguaglianza delle matrici di concentrazione e `alpha obs.2` descrive il livello di significatività osservato quando viene testata l'ipotesi di uguaglianza delle medie nei due gruppi sperimentali. In tutte le matrici su cui è stata condotta l'analisi è stata considerata significativa la *clique* che risultava essere significativa in almeno uno dei due test svolti. Nelle tabelle riportate qui sotto sono state riportate solo le *cliques*

[11] http://cran.r-project.org/web/packages/topologyGSA/index.html

considerate significative. Va ricordato che se una *clique* risulta significativa nel primo sistema d'ipotesi allora questo significa che la forza delle relazioni che la definiscono ha subito un cambiamento. Se, invece, una *clique* risulta essere significativa per il secondo sistema d'ipotesi allora ciò significa che è cambiata la media dell'espressione. Nella matrice "MI" sono risultate significative 12 *cliques* tra le 30 su cui sono state svolti i test.

CLIQUES SIGNIFICATIVE	alpha obs.1	alpha obs.2
CDK6, CCND1, CDK4, E2F1, E2F2, E2F3, RB1	0.0227	0.0050
CDKN1B	0.0453	0.8709
CDKN2A, MDM2, TP53	0.0347	0.0070
CTBP1, CTBP2, MECOM, HDAC1, HDAC2, RUNX1	0.0044	0.3090
MYC, BCR, ABL1	$6.196e^{-05}$	0
STAT5A, BCR, ABL1, STAT5B	$9.48e^{-05}$	0
GAB2, CRK, CRKL, CBL, CBLB, BCR, ABL1	0.0056	0
GAB2, GRB2, BCR, ABL1	0.0332	0
CHUK, IKBKB, NFKB1, NFKBIA, RELA	0.2890	0.0061
SHC2, BCR, ABL1	0.0189	0
SHC3, BCR, ABL1	0.0437	0
SHC1, BCR, ABL1	0.0027	0

Tabella 4.9: **Le cliques significative nella matrice "MI"**

Le *cliques* significative nelle matrici "ME" e "CP" sono 11, mentre quelle identificate nella matrice "MMDE" sono 13 e quelle identificate nella matrice "MDE" sono 14 tutte sul totale delle 27 *cliques* su cui stati svolti i test. Nelle tabelle dalla 4.10 alla 4.13 sono riportate le *cliques*, e i geni che le compongono, significative in ognuna delle matrici semplificate, e i relativi livelli di significatività osservata per i due sistemi d'ipotesi. In alcuni casi si può notare che le *cliques* significative sono composte, oltre che da geni anche da gruppi AND e OR e questo semplifica no-

tevolmente la lunghezza e la numerosità delle *cliques*. Ad esempio, nel caso della matrice "MI" era stata individuata come significativa la *clique* "CDK6, CCND1, CDK4, E2F1, E2F2, E2F3, RB1", composta quindi da 7 geni. Nelle matrici semplificate invece risulta significativa la *clique* "CCND1.CDK4, E2F, RB1" composta da un gene, un gruppo OR e un gruppo AND, e questa è esattamente la clique individuata all'inizio. In "MI", inoltre, erano risultate significative le cliques "SCH2, BCR", "SHC3, BCR", "SHC1, BCR", questo viene riassunto nelle matrici semplificate dalla significatività di un'unica *clique* "ABL, SHC" formata da 2 gruppi OR.

CLIQUES SIGNIFICATIVE	alpha obs.1	alpha obs.2
ABL, MYC	0.0300	0.0010
CCND1.CDK4, E2F, RB1	0.0150	0.0640
MDM2, CDKN1A, CDKN2A	0.0245	0.0130
MDM2, CDKN1A, TP53	0.0389	0.0080
ABL, GAB2, GRB2	0.6767	0.0022
ABL, PIK, BAD	0.2544	$8.72e^{-08}$
ABL, STAT, BAD	0.3547	$1.027e^{-06}$
ABL, CRKL.CBLB, PIK, GAB2	0.4256	$4.535e^{-04}$
ABL, SHC	0.3354	0.0020
NFKBIA, NFKB1, CHUK	0.3020	0.0079
ATK, MDM2, CDKN2A	0.1844	0.0282

Tabella 4.10: **Le cliques significative della matrice "ME"**

CLIQUES SIGNIFICATIVE	alpha obs.1	alpha obs.2
CCND1.CDK4, E2F, RB1	0.0321	0.018
ABL, STAT, BAD	0.0037	0
ABL, CRKL.CBLB, PIK, GAB2	0.1633	$2.85e^{-09}$
ABL, MYC	0.0061	0
MDM2, CDKN1A, TP53	0.0389	0.0080
MDM2, CDKN1A, CDKN2A	0.0245	0.0110
ATK, MDM2, CDKN2A	0.1534	0.0171
CCND1.CDK4,CDKN2A, CDKN1A	0.0964	0.0375
ABL, SHC	0.3423	$7.67e^{-11}$
NFKBIA, NFKB1, CHUK	0.4827	0.0057
ATK, PIK, BAD	0.0408	0.1640
ABL, PIK, BAD	0.0241	0
ABL, GRB2, GAB2	0.2194	$1.295e^{-10}$

Tabella 4.11: **Le cliques significative della matrice "MMDE"**

CLIQUES SIGNIFICATIVE	alpha obs.1	alpha obs.2
CCND1.CDK4, E2F, RB1	0.0397	0.016
ABL, STAT, BAD	0.0037	0
ABL, CRKL.CBLB, PIK, GAB2	0.1342	$1.55e^{-09}$
ABL, MYC	0.0061	0
MDM2, CDKN1A, TP53	0.0389	0.0060
MDM2, CDKN1A, CDKN2A	0.0245	0.0200
ATK, MDM2, CDKN2A	0.1534	0.0171
CCND1.CDK4,CDKN2A, CDKN1A	0.1330	0.0360
ABL, SHC	0.3423	$7.67e^{-11}$
ATK, PIK, BAD	0.0408	0.1640
ABL, PIK, BAD	0.0241	0
ABL, GRB2, GAB2	0.2194	$1.295e^{-10}$
TGFBR1.TGFBR2, SMAD3, SMAD4	0.4240	0.5160
NFKBIA, NFKB1, CHUK	0.4827	0.0057

Tabella 4.12: **Le cliques significative della matrice "MDE"**

CLIQUES SIGNIFICATIVE	alpha obs.1	alpha obs.2
CCND1.CDK4, E2F, RB1	0.3446	0.0211
ABL, STAT, BAD	0.6162	$1.645e^{-05}$
ABL, CRKL.CBLB, PIK, GAB2	0.9285	0.0096
ABL, MYC	0.0267	0.0050
MDM2, CDKN1A, TP53	0.0389	0.0090
MDM2, CDKN1A, CDKN2A	0.0245	0.0190
ATK, MDM2, CDKN2A	0.0862	0.0314
ABL, SHC	0.9706	0.0124
ABL, PIK, BAD	0.5723	$1.016e^{-05}$
ABL, GRB2, GAB2	0.4378	0.0179
NFKBIA, NFKB1, CHUK	0.4353	0.0161

Tabella 4.13: **Le cliques significative della matrice "CP"**

Nelle seguenti figure viene data una rappresentazione grafica delle posizioni che assumono i geni, che compongono le *cliques* significative, all'interno del *pathway* espanso (figura 4.3) e all'interno del *pathway* semplificato in ognuno delle quattro matrici (figure 4.4, 4.5, 4.6, 4.7). Come precedentemente detto le cliques vengono individuate a partire dal grafo moralizzato del *pathway*, di cui si può vedere una rappresentazione in Appendice (la figura 3 si riferisce al grafo moralizzato del *pathway* espanso, mentre la figura 4 si riferisce al grafo moralizzato del *pathway* semplificato). Le *cliques* significative vengono comunque rappresentate sui due *pathway* in modo da rendere più comprensibile una visione complessiva della situazione topologica.

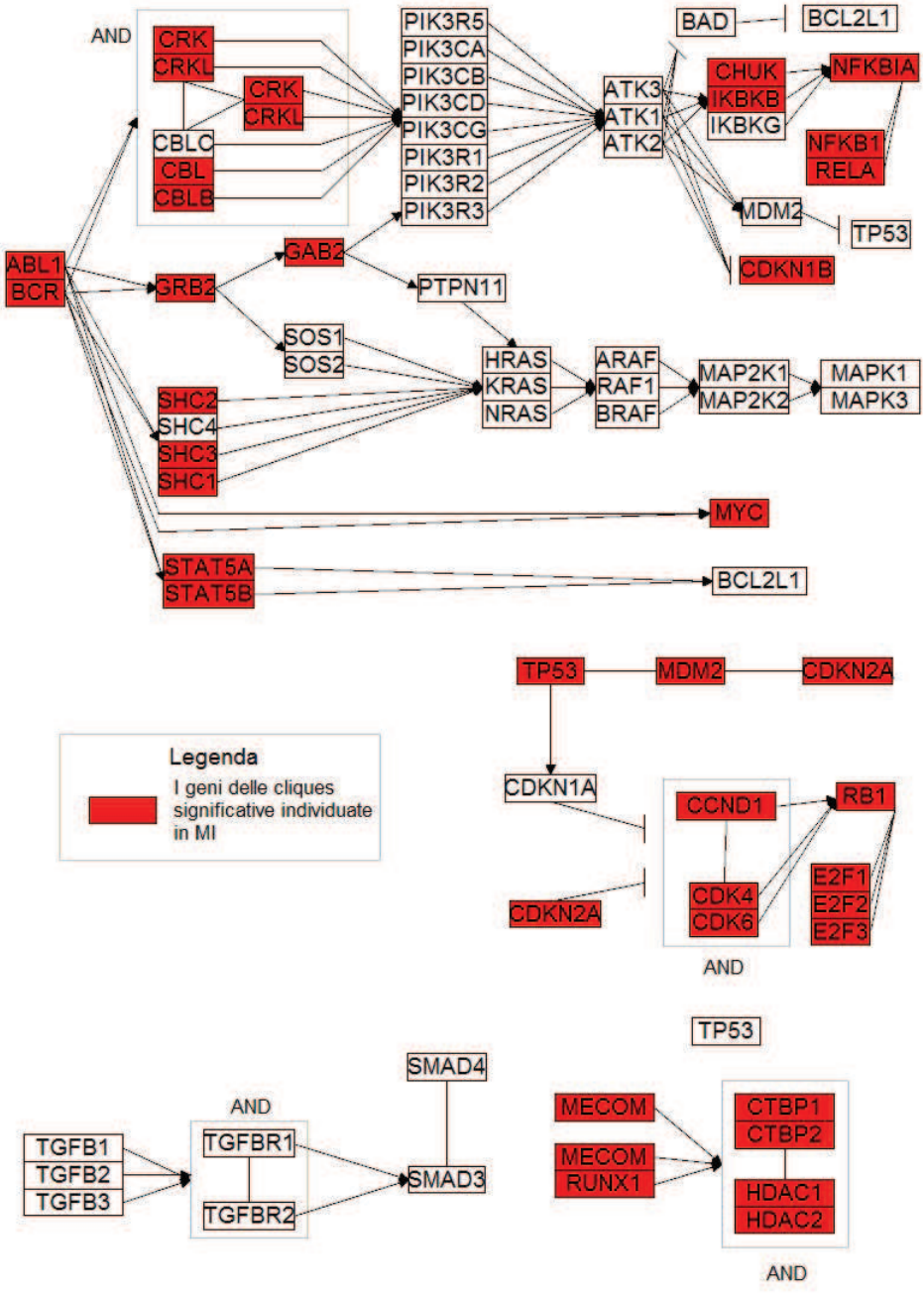

Figura 4.3: Le cliques significative del pathway espanso

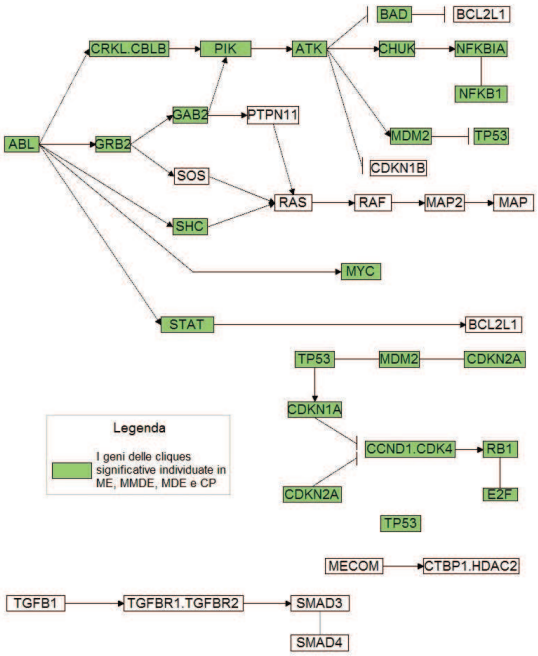

Figura 4.4:Le cliques significative della matrice "ME"

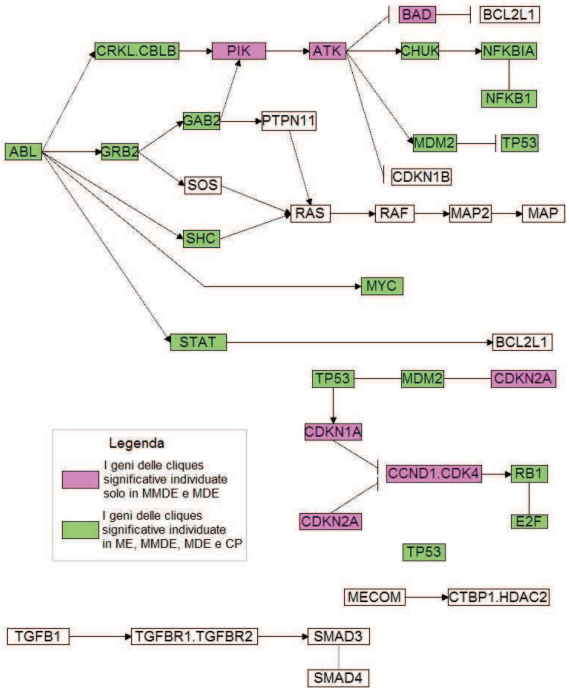

Figura 4.5: Le cliques significative nella matrice "MMDE"

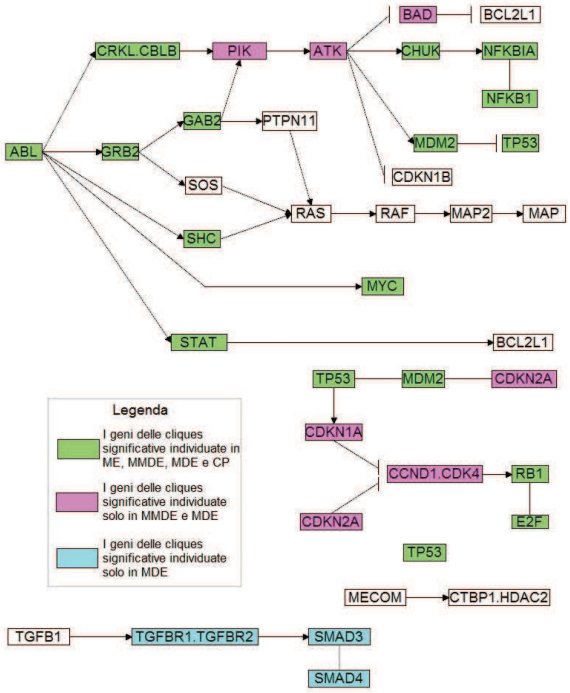

Figura 4.6: Le cliques significative nella matrice "MDE"

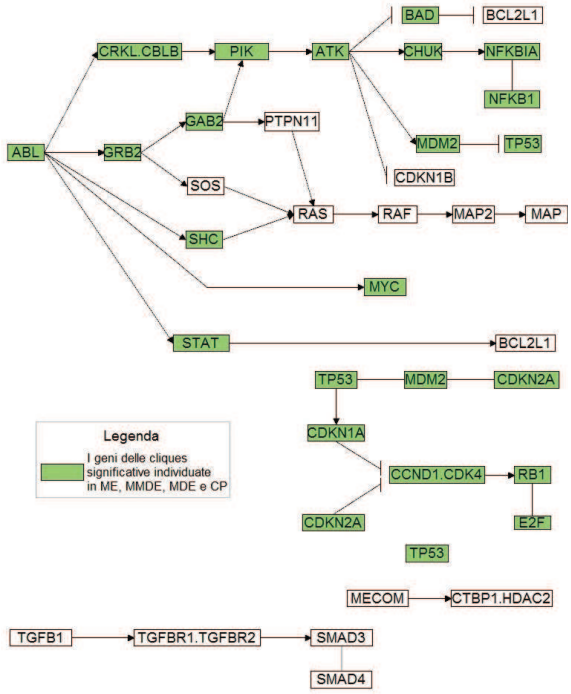

Figura 4.7: Le cliques significative nella matrice "CP"

Il grafico riportato in figura 4.8 mostra quante, tra le *cliques* significative trovate nelle matrici semplificate, sono state individuate dalle diverse situazioni considerate. Si vede come tutte le *cliques* identificate nelle matrici "ME" e "CP" siano state identificate anche in "MMDE" e "MDE". Le *cliques* "CCND1.CDK4, CDKN2A, CDKN1A" e "ATK, PIK, BAD" sono state identificate solo nelle matrici "MMDE" e "MDE". Mentre la *clique* "TGFBR1.TGFBR2, SMAD3, SMAD4" è stata identificata solo in "MDE".

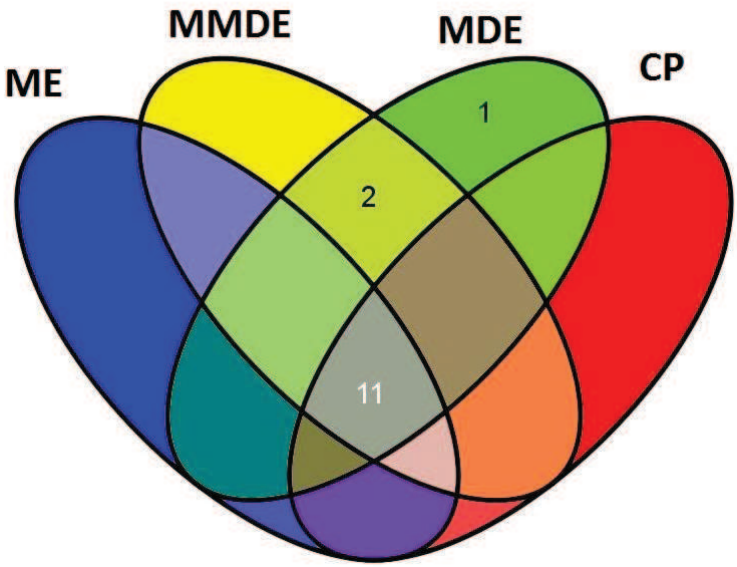

Figura 4.8: Diagramma di Venn: le cliques in comune nelle matrici semplificate

4.5 L'analisi SPIA

In questo paragrafo vengono esposti i risultati dell'analisi SPIA svolta su ognuna delle 5 situazioni già specificate precedentemente. Per ottenere questi risultati è stata utilizzata la funzione `runSPIA()` nella matrice in esame, opportunamente preparato attraverso la funzione `prepareSPIA()`; entrambe le funzioni fanno parte del pacchetto `graphite`[12]. Nel seguito sono riportati i risultati dell'analisi nella quale sono messe a confronto le statistiche di sintesi ottenute in tutti le matrici analizzate. Poiché la funzione `runSPIA()` richiede un vettore di *fold-changes*, espressi attraverso il logaritmo in base 2, dei geni identificati come differenzialmente espressi, si sono considerate due situazioni:

- Si sono considerati geni differenzialmente espressi quei geni a cui è associato un p.value corretto per la statistica test (2.19) inferiore a 0.10;
- Si sono considerati geni differenzialmente espressi quei geni a cui è associato un p.value non corretto per la statistica test (2.19) inferiore a 0.10;

Nella tabella 4.14, che fa riferimento alla prima delle due situazioni, sono riportate le principali statistiche di sintesi specificatamente per le 5 matrici. In particolare:

- pSize indica il numero di geni che compongono il *pathway*;

[12] http://www.bioconductor.org/packages/release/bioc/html/graphite.html

- NDE indica il numero di geni differenzialmente espressi del *pathway*;
- t_A è il valore osservato della perturbazione accumulata totale;
- P_{NDE} (3.9) che descrive la probabilità di osservare almeno un numero di DEG, pari a NDE, nel *pathway* utilizzando un modello ipergeometrico;
- P_{PERT} (3.12) che descrive la probabilità di osservare un valore più estremo di t_A;
- P_G (3.14) è il valore del p.value ottenuto dalla combinazione di P_{NDE} e P_{PERT}.

Come è stato detto precedentemente, il numero di geni che compone il *pathway* espanso è superiore rispetto al caso in cui si consideri il *pathway* semplificato. Nel caso in cui si considerino i log2 *fold-changes* dei geni DE a un livello di significatività corretto pari a 0.10 (tabella 4.14), l'analisi riscontra 2 geni differenzialmente espressi in tutte le casistiche tranne quella in cui si considera come sintesi "CP". La probabilità P_{NDE} è dello stesso ordine di grandezza (0.6436 e 0.6697 rispettivamente) se l'analisi viene svolta nelle matrici "MI" e "CP". E' identica, invece, e minore della precedente se si considerano le altre 3 matrici ($P_{NDE} = 0.298$). Questi risultati stanno a indicare che l'analisi di arricchimento migliora, rispetto alla situazione iniziale, utilizzando una qualsiasi delle sintesi "ME", "MMDE" o "MDE". Oltre al valore di P_{NDE}, le matrici "MMDE" e "MDE" mostrano gli stessi identici valori anche per tutte le altre

statistiche calcolate. Le analisi nella matrice "CP" appaiono subito non buone, ma questo è confermato anche dai valori di $t_A = 0$ e $P_{PERT} = 1$. Per quanto riguarda il valore di P_{PERT} si nota che nel caso di "MMDE" e "MDE" questo risulta inferiore (0.21) a quello trovato nell'analisi in "MI" (0.296), contrariamente a "ME" il cui valore di probabilità arriva a superarlo nettamente ($P_{PERT} = 0.678$). L'analisi topologica su "MMDE" e "MDE", quindi, risulta essere migliore sia rispetto a qualsiasi altra sintesi utilizzata sia rispetto alla situazione iniziale in cui tutti gruppi OR e AND risultano espansi. Questa tendenza è riscontrata anche dai valori di P_G.

	MI	ME	MMDE	MDE	CP
p Size	68	34	34	34	34
NDE	2	2	2	2	1
P_{NDE}	0.6436	0.2982	0.2985	0.2985	0.6697
t_A	-1.1709	-0.4633	-1.7332	-1.7332	0
P_{PERT}	0.2960	0.6780	0.2100	0.2100	1
P_G	0.5064	0.5257	0.2363	0.2363	0.9382

Tabella 4.14: **Risultati SPIA con DEG a un livello 0.1 del p.value adj.**

Dal momento che considerare come soglia per l'identificazione dei geni DE un p.value corretto risulta un approccio molto conservativo, si è deciso di ripetere l'analisi considerando come soglia per l'identificazione dei geni DEG un p.value non corretto pari sempre a 0.10. I risultati sono riportati nella tabella 4.15. In questo caso il numero di geni DE nella matrice "MI" è 14. In "MMDE" e "MD" vengono identificati 12 geni DE, mentre nelle altre matrici il risultato risulta meno buono, soprattutto in "ME" in cui sono identificati solo 7 geni DE. Il valore di P_{NDE}

diminuisce drasticamente in tutti i casi tranne in "ME", dove aumenta. Il valore di P_{PERT} aumenta notevolmente in "MI" (da 0.296 a 0.677), aumenta di poco in "MMDE" e in "MDE", mentre diminuisce negli altri casi. E' notevole soprattutto la diminuzione che ha avuto in "CP" dove la probabilità passa da 1 a 0.498. Infine il valore di P_G, rispetto all'analisi precedente, aumenta sia in "MI" che in "ME", mentre diminuisce in tutte le altre situazioni.

	MI	ME	MMDE	MDE	CP
p Size	68	34	34	34	34
NDE	14	7	12	12	8
P_{NDE}	0.4522	0.5026	0.0225	0.0225	0.3354
t_A	-0.7862	-0.6182	-1.9372	-1.9015	-0.8425
P_{PERT}	0.6770	0.6030	0.2770	0.2860	0.4980
P_G	0.6685	0.6649	0.0379	0.0389	0.4659

Tabella 4.15: **Risultati SPIA con DEG a un livello 0.1.**

5 Conclusioni

Nell'ambito delle analisi di esperimenti di *microarray* tra due gruppi di campioni i risultati che vengono spesso forniti consistono in una lista di geni differenzialmente espressi con le relative stime dei cambiamenti nei livelli di espressione nei due gruppi sperimentali. Questi geni possono essere associati tra loro da un punto di vista funzionale, come nel caso di un *pathway* biologico. Molti dei metodi utilizzati per la *pathway analysis* considerano il *pathway* come un'entità che fornisce solo una lista di geni funzionalmente connessi, ovvero un *gene set*. In realtà il *pathway* fornisce importanti informazioni anche sulla struttura della correlazione esistente tra i geni. Questo aspetto viene preso in considerazione da due metodologie, TopologyGSA e SPIA, che sono state utilizzate per svolgere le analisi in questo elaborato. In particolare, si è prestata particolare attenzione alle singole componenti del *pathway* in modo da poter identificare quali tra queste sono maggiormente coinvolte nel processo di sregolazione dell'intero *pathway*. Questo è stato possibile grazie alla complessa conversione del *pathway* in un modello grafico. Per poter arrivare a un modello grafico inizialmente è stato convertito il *pathway* in un grafo aciclico orientato (DAG) il quale è stato successivamente moralizzato e,

se ritenuto necessario, triangolarizzato. All'interno di un *pathway* molto spesso i nodi corrispondono a prodotti genici, ovvero complessi proteici (gruppi AND) o geni con funzioni biochimiche simili (gruppi OR). Il segnale propaga in modo differente tra questi due gruppi, quindi è molto importante considerarli in due modi diversi, ovvero i gruppi AND espansi in una *clique* mentre i gruppi OR espansi senza considerare le connessioni tra i geni che li compongono. Partendo da un *pathway* in cui sia i gruppi AND che i gruppi OR risultano espansi, lo scopo di questa tesi è stato quello di cercare e valutare un metodo che consenta di semplificare la struttura del *pathway*, in modo da rendere l'analisi computazionalmente più veloce, e che riesca a fornire delle stime precise riguardo l'identificazione dei sottogruppi genici sregolati. Nel 70-75% di pazienti affetti da leucemia linfoblastica acuta si è notata una differenziazione a livello delle cellule B, in particolare la traslocazione del gene ABL dal cromosoma 9 al cromosoma 22 con conseguente formazione di un gene chimera BCR/ABL. Il 95% dei paziente affetti da questo difetto cromosomico è affetto da un particolare tipo di leucemia, la leucemia mieloide cronica. Per questo motivo il *pathway* su cui ci siamo concentrati è il *pathway "Chronic Myeloid Leukemia"* che contiene tale gene chimera. Tra i metodi di semplificazione utilizzati quelli che saranno considerati migliori saranno quelli che forniranno alti livelli di significatività nell'identificazione dei geni differenzialmente espressi, nell'analisi topologica sul *pathway* e nell'analisi topologica sulle *cliques*. Partendo da una situazione iniziale in cui vengono identificati 68 geni come nodi del *pathway* "Chronic Myeloid Leukemia", si sono considerati 4 metodi come possibili alternative

per la sintesi del *pathway*, le quali sono composte da 34 nodi ciascuna. Successivamente sulla matrice dei dati iniziale e sulle 4 matrici semplificate sono state svolte le stesse analisi topologiche. Questo ci permette di confrontare i risultati ottenuti e individuare quali potrebbero essere le sintesi migliori per poter svolgere una semplificazione del *pathway*. Le matrici semplificate sono:

- "ME" in cui i gruppi AND e OR sono rappresentati attraverso la media aritmetica;
- "MMDE" in cui i geni che compongono i gruppi OR sono rappresentati dal gene che presenta una maggiore differenziale espressione in media tra il gruppo BCR e il gruppo NEG, mentre per i gruppi AND è stata considerata come sintesi la media aritmetica;
- "MDE" e in cui si prendono come rappresentanti dei gruppi OR e AND i geni che presentano una maggiore differenziale espressione in media tra il gruppo BCR e il gruppo NEG;
- "CP" in cui i livelli di espressione dei gruppi OR e dei gruppi AND vengono rappresentati dalla prima componente principale.

In tutte le matrici in esame sono state condotte le seguenti analisi:

- Identificazione dei geni differenzialmente espressi;
- Analisi topologica sul *pathway* completo usando i modelli grafici Gaussiani (topologyGSA);

- Analisi topologica sulle *cliques* usando i modelli grafici Gaussiani (topologyGSA);
- Analisi topologica con SPIA.

Nell'analisi di identificazione dei geni differenzialmente espressi la situazione che offre migliori risultati è quella che considera come sintesi dei livelli di espressione per i gruppi OR la massima differenziale espressione in media, indipendentemente da come vengono sintetizzati i gruppi AND. Questa metodologia riesce a evidenziare anche la differenziale espressione dei gruppi AND e OR che contengono geni identificati come DE nella matrice "MI". Le altre tipologie di sintesi non danno risultati altrettanto buoni, non identificano un alto numero di geni sregolati e soprattutto non riescono a evidenziare la differenziale espressione dei gruppi AND e OR.

Nell'analisi topologica del *pathway* tutte le metodologie considerate, tranne "CP", riescono a cogliere la sregolatezza del *pathway* dovuta sia alla correlazione tra i geni che alla media dell'espressione nelle due condizioni sperimentali.

La stessa analisi svolta sulle *cliques* ha evidenziato che tutte le *cliques* identificate da "ME" e "CP" vengono identificate da "MMDE" e "MDE". Queste ultime due hanno in comune le *cliques* "CCND1.CDK4, CDKN2A, CDKN1A" e "ATK, PIK, BAD". Solo in "MDE" risulta significativa "TGFBR1.TGFBR2, SMAD3, SMAD4".

Nell'analisi SPIA, considerando il caso in cui i geni DEG vengono selezionati a una soglia del p.value non corretto di 0.10 in modo da avere

dei risultati che mostrino in modo più marcato le differenze tra i metodi di sintesi, viene ribadito il fatto che le sintesi ottenute con le medie e le componenti principali non danno buoni risultati. Anche in questo caso le metodologie che offrono risultati migliori sono quelle in cui i livelli di espressione dei gruppi OR vengono sintetizzati dal gene che presenta massima differenziale espressione in media.

Per concludere, dopo aver svolto tutte le analisi esposte in questo elaborato e dopo aver confrontato i risultati, sembra che utilizzare come sintesi per i gruppi AND e OR la prima componente principale e la media aritmetica non siano i modi migliori per semplificare la situazione iniziale corrispondente al *pathway* espanso. Un buon risultato invece è dato dalla sintesi che utilizza per i gruppi OR la massima differenziale espressione, anche se non si può affermare con assoluta sicurezza che utilizzare come sintesi dei livelli di espressione per i gruppi AND la media aritmetica piuttosto che la massima differenziale espressione dia dei risultati migliori o viceversa.

Nonostante i metodi di semplificazione testati in questo elaborato abbiamo dato dei buoni risultati su un singolo *pathway,* per valutarne l'affidabilità sarebbe opportuno fare un'analisi completa su tutti i *pathway* (come generalmente si fa nelle analisi *genome-wide*) magari con un numero maggiore di gruppi AND e OR. Sarebbe, inoltre, interessante applicare questi metodi di sintesi alle analisi in cui vengono considerate le

interazioni esistenti tra più *pathway*. Purtoppo recuperare le informazioni riguardo la composizione dei gruppi non è a tutt'oggi un'operazione automatica, in questo elaborato è stato fatto manualmente attraverso il formato KGML. Sarebbe utile sviluppare un *software* che sia in grado di automatizzare tale processo.

Appendice

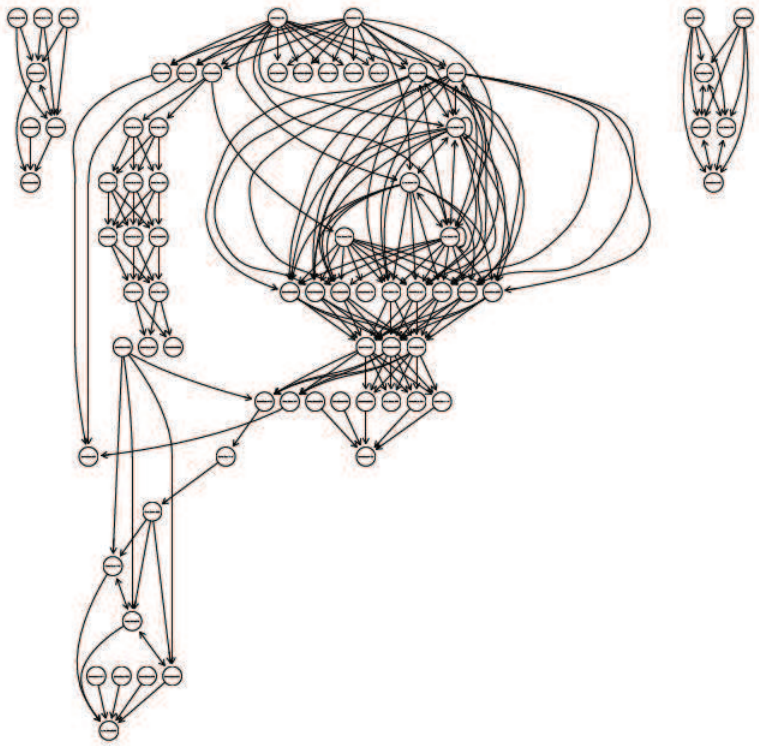

Figura 1: Il grafo espanso KEGG

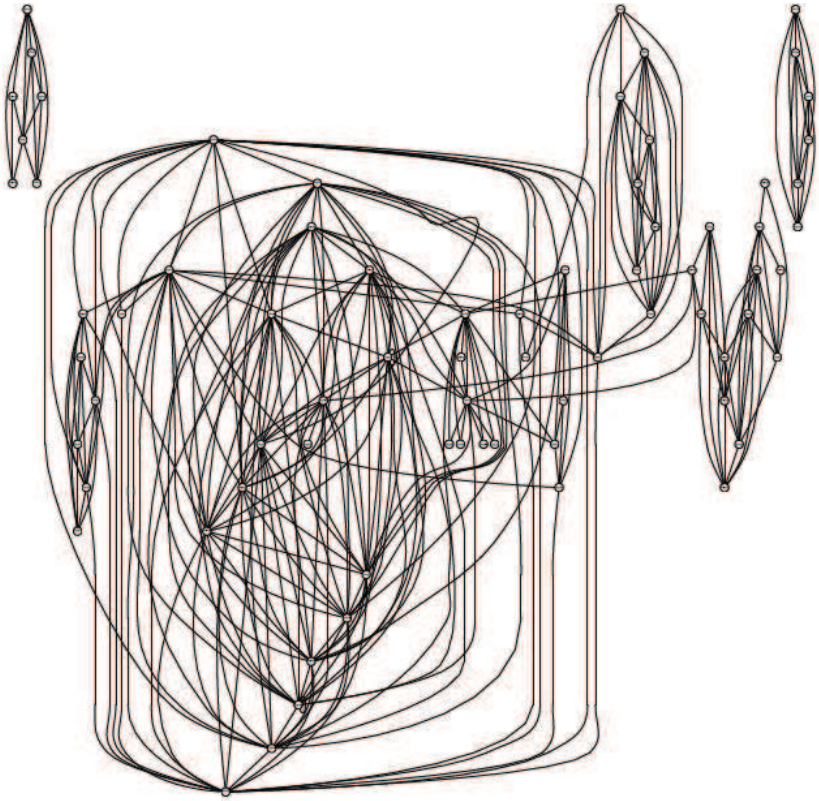

Figura 2: Grafo moralizzato Kegg

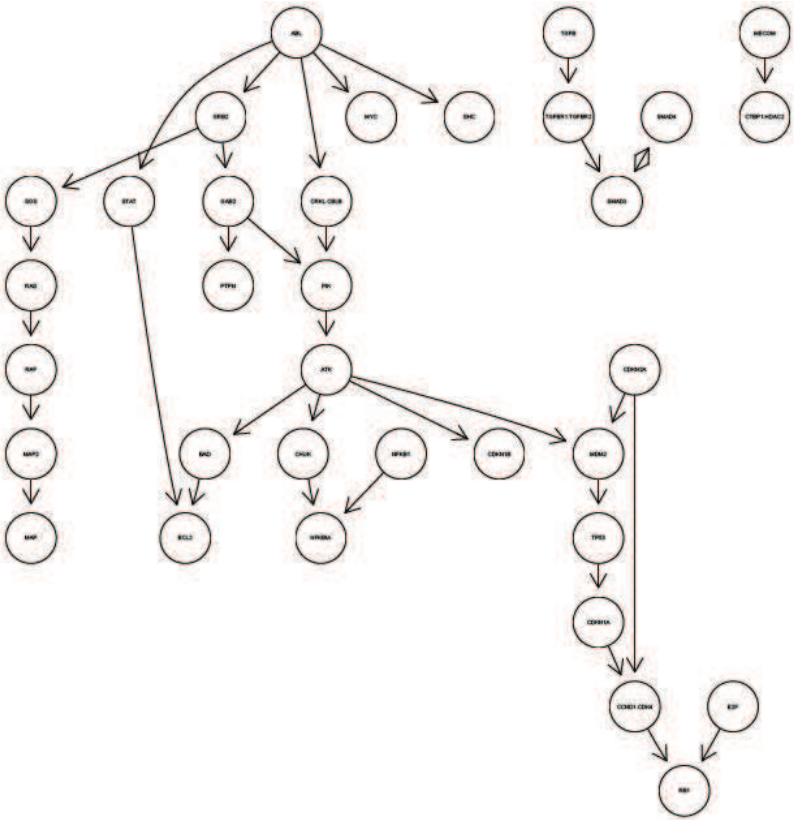

Figura 3: Il grafo del pathway semplificato

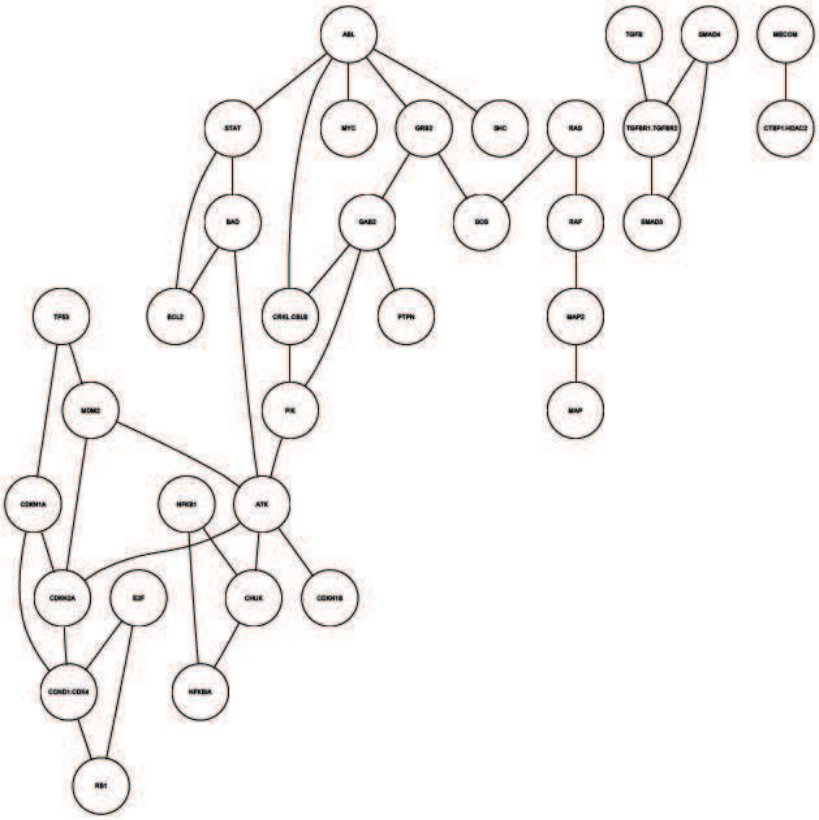

Figura 4: Il grafo moralizzato del pathway semplificato

Bibliografia

[1] Smyth, G. K. **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Statistical Applications in Genetics and Molecular Biology*, **3**, No. 1, Article 3. (2004) http://www.bepress.com/sagmb/vol3/iss1/art3.

[2] Smyth, G. K. **Limma: linear models for microarray data User's Guide.** (2003) Software manual available from http://www.bioconductor.org.

[3] B.M.Bolstad ,R.A.Irizarry, M.A. Strand and T.P.Speed. **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias** *Bioinformatics*, **19**: 185-193, 2003.

[4] R.A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P.Speed, **Exploration, normalization, and summaries of high density oligonucleotide array probe level data** *Biostatistics*, 4: 249-264, (2003).

[5] Kerr, M. K., and Churchill, G. A. **Experimental design for gene expression microarrays.** *Biostatistics 2*, 183-201. (2001)

[6] Draghici s, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. **A system biology approch for pathway level analysis.** *Genome Research*, 17(10):1537-1545 (2007).

[7] Massa M, Chiognia M, Romualdi C. **Gene set analysis exploiting tha topology of a pathway.** *BMC Systems Biology*, 4: 121. (2010).

[8] G.Sales, E.Calura, D.Cavalieri, C.Romualdi. **Graphite - a Bioconductor package to convert pathway topology to gene network**, *BMC Bioinformatics*, 13: 20 (2012).

[9] Goldman JM,Melo JV. **Chronic myeloid leukemia**, *advances in biology and new approaches to treatment.* N Engl J Med; 349:1451-1464. (2003)

[10] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids* Res.27(1):29-34. (1999).

[11] S. Chiaretti, X. Li, R. Gentleman, A.Vitale, K.S. Wang, F.Mandelli, R.Foà, J.Ritz, **Gene expression profiles of B-lineage adult Acute Lymphocytic Leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of trasformation**, Clinical Cancer Research (2005);11(20):7209-19

[12] Lauritzen SL: **Graphical models**, *Clarendon Press*, Oxford (1996)

[13] Tsai CA, Chen JJ: **Multivariate analysis of variance test for gene set analysis**. *Bioinformatics* 2009, 25: 897-903.

[14] Sales G, Calura E, Romualdi C: GRAPH Interaction from pathway Topological Environment.

[15] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J, Kim CJ, Kusanovic JP, Romero R,: **A novel Signaling Pathway Impact Analysis (SPIA)**, Bioinformatics (2008).

[16] Benjamini Y, Hochberg Y, **Controlling the false discovery rate: a pratical and powerful approach to multiple testing**, *Journal of the Royal Statistical Society* (1995).