UNIVERSITÀ DI PADOVA  FACOLTÀ DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea Magistrale in Bioingegneria

Tesi di Laurea

ANALISI DI CORRELAZIONE TRA GENI PER LA SELEZIONE ROBUSTA DI BIOMARCATORI

Relatore: Dott.ssa Barbara Di Camillo

Correlatori: Dott.ssa Tiziana Sanavia

Dott. Marco Mina

Laureanda: Roberta Mazzucco

Anno accademico 2010/2011

Padova, 25 Ottobre 2011

A mamma, papà,

Richi e Faz.

Indice

SOMMARIO

Nell'ultimo decennio gli studi sul trascrittoma effettuati con i microarray hanno avuto una massiccia diffusione, consentendo di disporre di dati di espressione genica che si sono dimostrati estremamente utili nell'identificare geni biomarcatori di malattie genetiche complesse quali i tumori. Tuttavia, le tecniche comunemente adottate per la selezione dei biomarcatori tengono conto di variazioni nel livello di espressione genica, ignorando le relazioni di interazione tra geni e la loro alterazione indotta dalla malattia. In questo lavoro viene introdotto un nuovo metodo per il ranking dei geni che integra informazioni sulle interazioni geniche e si basa sulla misura delle variazioni di interazione di ciascun gene con gli altri, nell'ipotesi che le più significative siano quelle a carico dei biomarcatori. In particolare vengono proposte quattro diverse misure: una basata sulla massima variazione del grado di correlazione tra coppie di geni, le altre sul confronto di proprietà topologiche locali delle reti di interazioni geniche costruite a partire dai dati. Le prestazioni delle diverse misure, valutate mediante dati simulati, sono state confrontate con quelle del test SAM per la selezione dei geni differenzialmente espressi. I risultati ottenuti costituiscono un'analisi preliminare per lo sviluppo di nuove metodologie di identificazione dei geni biomarcatori.

1

Introduzione

La diffusione, cui si è assistito negli ultimi 15-20 anni, delle tecnologie high-throughput per l'analisi del trascrittoma, quali ad esempio i DNA microarray [1, 2], ha profondamente modificato l'approccio allo studio di malattie genetiche complesse quali ad esempio i tumori o le malattie neurodegenerative. La misura del livello di espressione dei singoli geni, mediante quantificazione delle diverse molecole di mRNA cellulare, consente di indagare i complessi meccanismi di controllo dell'espressione genica e di analizzare le possibili modificazioni dell'attività regolatoria che, indotte dalla malattia, danno luogo all'alterazione dei processi biologici e delle funzioni fisiologiche cellulari. La vera rivoluzione introdotta dai microarray è data dalla possibilità di monitorare contemporaneamente il livello di espressione degli oltre 20000 geni umani conosciuti su una stessa piattaforma, resa possibile dal crescente grado di miniaturizzazione consentito dalle nuove tecnologie. Il monitoraggio dell'intero trascrittoma, senza dover restringere l'analisi a singoli geni o pathway isolati, e la possibilità di considerare il sistema di regolazione genica nella sua globalità consente di studiare malattie complesse caratterizzate dalla presenza di numerose mutazioni e alterazioni funzionali a carico di diversi geni.

In particolare, l'analisi dei dati ottenuti mediante microarray si è dimostrata estremamente utile per l'identificazione di quei geni, detti biomarcatori, in grado di determinare lo stato patologico di un soggetto malato e di caratterizzare il manifestarsi e l'evolvere di una particolare patologia e la cui conoscenza è di fondamentale importanza nell'applicazione clinica per determinare il tipo di terapia. La diagnosi precoce, l'applicazione di una terapia specifica, la possibilità di prevedere l'evoluzione della malattia e la risposta dei pazienti ai trattamenti e ai farmaci somministrati e lo sviluppo di terapie geniche mirate sono solo alcune delle importanti conseguenze nella pratica clinica legate alla conoscenza dei geni biomarcatori [3, 4].

Nonostante la grande quantità di dati a disposizione, l'identificazione dei biomar-

catori resta un problema molto complesso: i metodi che si basano sul confronto dei dati di espressione provenienti da diversi esperimenti con microarray sullo stesso caso biologico producono spesso risultati differenti, mettendo in discussione l'affidabilità e la significatività delle liste di biomarcatori ottenute [5]. Le difficoltà nella selezione dei geni biomarcatori e la scarsa riproducibilità dei risultati sono riconducibili a caratteristiche intrinseche dei dati di espressione genica e della tecnologia dei microarray:

- il numero di trascritti monitorati è dell'ordine delle migliaia o delle decine di migliaia: se da un lato la possibilità di monitorare l'intero trascrittoma consente di considerare il sistema nella sua globalità, dall'altro complica la ricerca delle variabili che caratterizzano la patologia analizzata;

- in genere è possibile monitorare l'intero trascrittoma (migliaia o decine di migliaia di trascritti) a fronte di qualche decina o al massimo centinaia di array, sia per motivi tecnici ed etici legati alla raccolta dei campioni che per ragioni di costo: il numero di campioni disponibili per ciascuna variabile genica è di gran lunga inferiore al numero di variabili monitorate e il problema risulta essere mal condizionato [6];

- la variabilità tecnica e biologica che caratterizza gli esperimenti con microarray introduce una componente di rumore ai dati di espressione genica che ne complica l'analisi e l'elaborazione [7];

- i laboratori di analisi utilizzano piattaforme tecnologiche diverse e protocolli sperimentali e di elaborazione dei campioni biologici difficilmente riproducibili [8, 9];

- patologie complesse, quali i tumori e le malattie neurodegenerative, presentano un certo grado di eterogeneità poiché sono caratterizzate dall'alterazione di diversi pathway di regolazione piuttosto che da disfunzioni a carico di un singolo gene come per le malattie monogeniche [10].

In un tipico disegno sperimentale i dati di espressione provengono da soggetti diversi con fenotipi diversi che possono essere, ad esempio, soggetti sani e soggetti

malati o soggetti che presentano la stessa malattia ma in forme diverse o ad un diverso stadio evolutivo. Dal confronto dei profili di espressione genica di due diverse classi di popolazione è possibile individuare i geni la cui attività è stata alterata dalla malattia e che risultano sovraregolati o sottoregolati nei soggetti malati rispetto a quelli di controllo e ordinare i geni in base alla loro capacità di discriminare i soggetti delle due classi.

Le tecniche più frequentemente adottate per l'identificazione dei biomarcatori a partire da dati di microarray si basano sull'utilizzo di test statistici univariati (parametrici o non parametrici) per la selezione dei geni differenzialmente espressi, ovvero con livello di espressione significativamente diverso nelle due popolazioni di soggetti messe a confronto dal punto di vista statistico [11–13]. Il limite principale di questo approccio è quello di considerare le variabili singolarmente, ignorando completamente l'esistenza di possibili relazioni tra geni. L'utilizzo di test statistici multivariati per l'identificazione dei geni differenzialmente espressi [14, 15] o, in un contesto di classificazione, per la selezione delle variabili che meglio discriminano i soggetti delle due classi massimizzando le prestazioni di un classificatore [16], consente di superare in parte questa limitazione. Nonostante i metodi multivariati tengano conto di come sottoinsiemi di variabili contribuiscano tra loro per spiegare i dati, nessuno dei metodi comunemente utilizzati per la selezione dei biomarcatori tiene esplicitamente in considerazione le vere e proprie relazioni di interazione tra geni.

Le alterazioni dell'attività genica e dei complessi meccanismi di controllo dell'espressione indotte dalle malattie genetiche si ripercuotono sulle relazioni di interazione tra geni che stanno alla base della rete di regolazione. La possibilità di tenere in considerazione tali relazioni può essere dunque rilevante ai fini dell'identificazione dei geni biomarcatori che caratterizzano il manifestarsi della patologia analizzata. Solo recentemente sono stati presentati alcuni lavori che prevedono l'integrazione di informazioni sulle interazioni geniche per la selezione dei biomarcatori: tali informazioni vengono anch'esse ricavate a partire dai dati di microarray mediante analisi del grado di correlazione tra variabili geniche. Tibshirani e Wasserman hanno proposto, in [17], un approccio per l'identificazione dei geni differenzialmente espressi detto "correlation-sharing": lo score assegnato a ciascun gene x si ottiene mediando i t-score (calcolati

mediante test statistico univariato SAM [13]) dei geni che presentano correlazione col gene x superiore ad una certa soglia determinata empiricamente a partire dai dati. Un altro metodo, proposto da Storey *et al.* in [18], è detto "Optimal Discovery Procedure" e prevede la definizione di una nuova statistica per la selezione dei geni differenzialmente espressi: il punteggio assegnato alla variabile x tiene in considerazione come la correlazione di x con le restanti variabili in gioco influenza il suo livello di espressione. Altri esempi di approcci basati sullo stesso principio sono il metodo proposto da Zuber e Strimmer [19] basato sulla definizione di un punteggio detto "correlation-adjusted t-score", o "cat-score", e i metodi proposti da Lai in [20] e da Zhang in [21] che tengono in considerazione i geni contemporaneamente espressi, o co-espressi, nel processo biologico o patologico analizzato.

Tenendo in considerazione i promettenti risultati presentati nei lavori appena citati, in questo lavoro di tesi vengono proposti dei nuovi metodi per il ranking dei geni e l'identificazione dei biomarcatori basati sull'integrazione di informazioni sul grado di correlazione tra variabili geniche. L'obiettivo è quello di migliorare le prestazioni dei metodi comunemente utilizzati, sia in termini di precisione che in termini di riproducibilità delle liste ordinate di geni ottenute. Pur basandosi sullo stesso criterio i nuovi metodi proposti si distinguono da quelli già citati per quanto concerne le assunzioni di base. I metodi presentati in [17–21] analizzano il grado di correlazione tra variabili geniche nell'ipotesi che i geni che caratterizzano la stessa patologia siano funzionalmente correlati, partecipino agli stessi pathway regolatori e siano co-espressi all'interno della cellula. Il nuovo metodo proposto, invece, si basa sull'assunzione, condivisa da medici e biologi, che i geni colpiti dalla malattia siano caratterizzati da una variazione significativa del livello di correlazione con gli altri geni.

Per la ricerca dei geni che hanno modificato in modo consistente le loro interazioni con gli altri geni, il nuovo metodo proposto si basa sul confronto delle reti di interazioni geniche costruite a partire dai dati di microarray di due classi di popolazione, mediante confronto pair-wise dei profili di espressione. In particolare, per ciascun gene vengono messe a confronto le proprietà topologiche locali in un intorno dei nodi ad esso associati nelle due reti costruite. Tuttavia, le caratteristiche intrinseche dei dati di espressione

genica di microarray, cui si è già fatto cenno, complicano notevolmente il problema di costruzione delle reti di interazioni geniche. Per poter individuare le variazioni di interazione significative dovute ad alterazioni dell'attività genica e che interessano i biomarcatori è dunque necessario utilizzare una misura di confronto della topologia che sia robusta al rumore presente nei dati ma soprattutto all'elevato numero di relazioni di tipo falso positivo inevitabilmente introdotte durante la costruzione delle reti. Per questo motivo è stata introdotta una misura che considera proprietà topologiche locali complesse e, in particolare, mette a confronto la partecipazione dei geni a strutture modulari anche molto articolate, dette *graphlets* [22]. Oltre a questa sono state proposte altre tre misure per la quantificazione delle variazioni di interazioni geniche, due delle quali basate sul confronto di proprietà topologiche più semplici, come la connettività e il coefficiente di clustering dei nodi, e la terza basata sul confronto diretto del grado di correlazione tra coppie di geni.

Per valutare le prestazioni del nuovo metodo e delle diverse misure proposte per il ranking dei geni in termini di precisione e recall è necessario conoscere la lista vera dei biomarcatori da identificare. A questo scopo sono stati simulati, mediante il simulatore descritto in [23], i dati di espressione genica di due popolazioni di 500 soggetti ciascuna e che presentano due forme diverse della stessa malattia. Per una valutazione della riproducibilità dei risultati, della stabilità delle liste ottenute e della robustezza al diminuire del numero di campioni di espressione per ciascun gene monitorato, le diverse misure sono state applicate a dataset di 50 e 25 soggetti ottenuti partizionando opportunamente quello iniziale. I risultati sono stati confrontati con quelli ottenuti mediante test statistico SAM per la selezione dei geni differenzialmente espressi [13].

Struttura della tesi

Nel *capitolo 2* viene descritto il procedimento adottato per la costruzione delle reti di interazioni geniche basato sul confronto pair-wise dei profili di espressione utilizzando come misure di similarità la correlazione di Pearson e la correlazione di Pearson condizionata, o correlazione parziale.

Il *capitolo 3* è dedicato all'introduzione del nuovo metodo per il ranking dei geni basato sull'integrazione di informazioni sul grado di correlazione tra geni; in particolare vengono descritte le diverse misure proposte per quantificare le variazioni di interazioni geniche.

Nel *capitolo 4* vengono descritte le fasi di simulazione per la generazione dei dati di espressione genica con particolare riferimento alla simulazione della variabilità biologica, dell'eterogeneità delle due forme della malattia riprodotte e del rumore di cui sono tipicamente affetti i dati di microarray.

Infine, nel *capitolo 5*, dopo una breve analisi delle proprietà topologiche delle reti di interazioni geniche costruite, vengono presentati e commentati i risultati ottenuti applicando ai dati simulati i nuovi metodi di ranking, sia in termini di precisione/recall che in termini di stabilità delle liste, confrontandoli con quelli ottenuti mediante test SAM.

In *appendice A* sono riportati i dettagli del metodo di correzione per test multipli basato su FDR, adottato nella costruzione delle reti per la selezione delle relazioni geniche significative, mentre in *appendice B* viene descritto il metodo di selezione dei geni differenzialmente espressi basato su test SAM.

2

Analisi della correlazione tra profili di espressione genica

La prima fase del lavoro svolto consiste nella costruzione di reti di interazioni tra geni: a partire dai dati di espressione genica di due diverse classi di popolazione di soggetti vengono ricostruite due reti, che, nella seconda fase del lavoro, verranno messe a confronto con l'obiettivo di individuare i geni biomarcatori. In questo capitolo, dopo una breve introduzione del contesto biologico in cui si opera, viene descritto il procedimento utilizzato per l'individuazione delle relazioni di correlazione tra geni, che si basa su metodi proposti per la *reverse engineering* delle reti di regolazione genica, in particolare sul confronto *pair-wise* dei profili di espressione genica. Vengono introdotte nel dettaglio le misure di correlazione utilizzate per confrontare i geni e vengono presentati i test statistici adottati per la selezione delle relazioni significative.

2.1 Reverse engineering delle reti di regolazione genica

Il termine *reverse engineering* indica l'insieme dei metodi impiegati per la ricostruzione della rete di regolazione e controllo a partire dall'output dinamico del sistema osservato. Il problema della reverse engineering delle reti di regolazione genica è uno dei problemi più interessanti nel campo della *Systems Biology* [24–26]: negli ultimi anni sono stati compiuti notevoli passi avanti e numerosissimi lavori sono stati pubblicati sull'argomento, tuttavia, data la complessità del problema, ancora oggi molto poco è noto circa la struttura delle reti di regolazione.

Tutte le cellule di un organismo contengono l'intera sequenza di DNA, all'interno della quale è codificato l'insieme delle istruzioni necessarie per produrre le proteine, molecole indispensabili alla vita dell'organismo e in grado di svolgere la maggior parte delle funzioni biologiche cellulari. La capacità delle cellule di assumere caratteristiche e ruoli specifici e differenziati è legata ad un differenziamento dell'espressione proteica: meccanismi di regolazione e controllo molto sofisticati e complessi consentono alle

cellule di produrre diverse proteine in diversi istanti temporali in base alle necessità e in risposta agli stimoli ricevuti. La regolazione dell'espressione genica avviene in più punti della cascata di reazioni che a partire dalle porzioni codificanti del genoma, i geni, porta alla produzione delle proteine: per attuare una raffinata regolazione del livello di espressione genica della cellula, la sequenza di DNA, le molecole di RNA messaggero e le proteine interagiscono tra loro formando una complessa rete di regolazione caratterizzata da diversi meccanismi di controllo. Oltre alla regolazione a livello della trascrizione del DNA in RNA messaggero, attraverso l'intervento di proteine regolatorie che agiscono attivando/inibendo o accelerando/rallentando il processo trascrizionale, esistono numerosi altri meccanismi di controllo. Un esempio sono il processo di *splicing* e di trasporto fuori dal nucleo della molecola di mRNA, come anche la sua eventuale degradazione, che sono regolati da diversi tipi di proteine. Inoltre proteine regolatorie intervengono durante la fase di traduzione della molecola di mRNA in proteina mediando il riconoscimento dell'mRNA da parte dei ribosomi e regolando l'efficienza e la velocità del processo di traduzione. Infine esiste un sistema di controllo che regola le interazioni proteina-proteina e le modifiche post-traduzionali le quali possono modificare la funzione di una proteina e la sua capacità di interagire con altre proteine e molecole e, quindi, la sua capacità di intervenire nella regolazione della trascrizione di altri geni.

L'avvento delle tecnologie *high-throughput* ha modificato in maniera consistente lo scenario della descrizione dei processi di regolazione: la conoscenza dell'intera sequenza genomica e la possibilità di monitorare l'intero trascrittoma o traduttoma di un organismo in una determinata condizione fisiologica, consentono di considerare il sistema nel suo complesso e di approfondirne la conoscenza in termini di meccanismi globali di regolazione senza dover restringere l'analisi ad un limitato numero di interazioni presenti nella rete di regolazione. L'analisi della struttura della rete e delle interazioni tra le molecole che prendono parte alla regolazione consente quindi di comprendere a fondo i processi biologici e fisiologici che avvengono all'interno della cellula e i meccanismi molecolari che caratterizzano alcune patologie, come ad esempio quelli che stanno alla base della trasformazione e progressione delle cellule tumorali. La

diffusione delle tecnologie high-throughput ha reso necessaria l'introduzione di nuovi modelli di reverse engineering adatti a trattare la grande mole di trascritti monitorati.

Nonostante la grande quantità di dati a disposizione, la ricostruzione delle reti di regolazione resta un problema estremamente complesso e caratterizzato da alcune limitazioni intrinseche [27]. In primo luogo le tecnologie high-throughput non consentono di monitorare tutti i diversi tipi di molecole che intervengono nel processo di controllo e le misure disponibili per la ricostruzione della rete di regolazione sono spesso limitate alle misure del trascrittoma tramite esperimenti con *microarray* [1, 2]. Le tecniche di monitoraggio del traduttoma, come ad esempio la spettrometria di massa, sono ancora in fase di sviluppo e di difficile applicazione e non sono disponibili misure che diano informazioni circa le modificazioni post-trascrizionali subite dalle molecole di mRNA e circa la loro influenza sui meccanismi regolatori. Di conseguenza i complessi meccanismi di regolazione, che prevedono l'interazione di geni, molecole di RNA di diverso tipo e proteine, possono essere ricostruiti soltanto facendo riferimento ad un modello semplificato, come rappresentato in figura 2.1: la quantità di mRNA viene considerata come un'approssimazione del livello di espressione proteica delle proteine da esso stesso codificate.

In secondo luogo le variabili che entrano in gioco sono estremamente numerose: il numero dei trascritti monitorati è dell'ordine delle migliaia o delle decine di migliaia. Se da un lato la possibilità di monitorare l'intero trascrittoma consente di considerare il sistema nella sua globalità, dall'altro dover risolvere sistemi caratterizzati da un numero così elevato di variabili complica notevolmente il problema, sia per i tempi computazionali elevati che per la difficoltà di individuare le variabili caratterizzanti il processo biologico di interesse tra le migliaia di geni monitorati, soprattutto quando non si dispone di alcuna conoscenza a priori circa il processo considerato.

Un'ulteriore complicazione al problema di ricostruzione della rete di regolazione è legata al fatto che tipicamente il numero di campioni disponibili per ciascuna variabile è di gran lunga inferiore al numero di variabili monitorate e il problema risulta essere mal condizionato. In genere è possibile monitorare l'intero genoma (migliaia o decine di migliaia di trascritti) a fronte di qualche decina o centinaia di array (sia per motivi

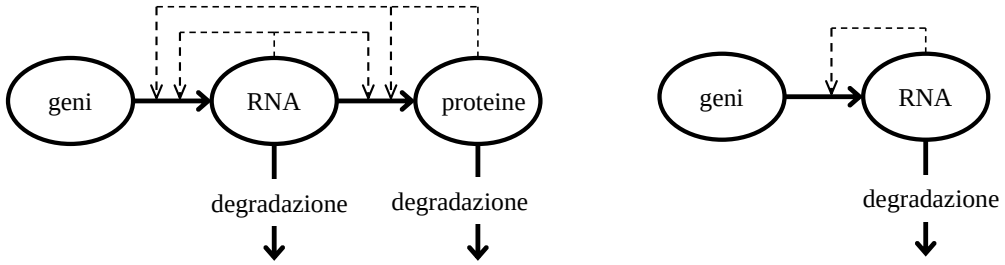


Figura 2.1 Rappresentazione schematica delle interazioni tra molecole che caratterizzano i meccanismi di regolazione (a sinistra) e modello semplificato in cui l'mRNA viene considerato come un'approssimazione della concentrazione proteica (a destra).

tecnici ed etici legati alla raccolta dei campioni, che per ragioni di costo).

Un ultimo fattore che condiziona la ricostruzione della rete è dato dal rumore di cui i dati di espressione genica sono affetti e dalla variabilità tecnica e biologica che caratterizza gli esperimenti con microarray: è dunque di fondamentale importanza utilizzare dei metodi di reverse engineering robusti al rumore e capaci di limitarne gli effetti sulla ricostruzione della rete.

La rete di regolazione genica ricostruita a partire da dati di microarray seguendo lo schema in figura 2.1 può essere rappresentata come un grafo diretto in cui i nodi identificano i geni e le proteine da essi codificate, mentre gli archi orientati rappresentano l'azione regolatoria del gene da cui l'arco si diparte sul gene in cui l'arco arriva. Negli ultimi anni sono stati proposti numerosi metodi di reverse engineering per inferire sulla rete di regolazione a partire dai dati dinamici di espressione genica, i quali possono essere suddivisi in due categorie principali: metodi basati su modello e metodi basati sul confronto pair-wise dei profili di espressione genica.

I primi definiscono un modello di regolazione e utilizzano i dati per identificarne i parametri o per massimizzare una funzione obiettivo che consenta di scegliere tra configurazioni alternative del modello quella che meglio descrive il sistema in analisi. Vi sono svariate tipologie di modelli che possono essere utilizzati per descrivere la rete da ricostruire, ciascuno dei quali è in grado di rappresentare specifiche proprietà delle reti di regolazione reali. Un primo esempio sono i modelli Booleani [28, 29] ben adatti a descrivere alcuni aspetti legati all'interazione tra attivatori e inibitori dei meccanismi

di trascrizione genica. I modelli Bayesiani [30, 31], tra i più diffusamente utilizzati in letteratura, sono in grado di descrivere la natura intrinsecamente probabilistica della regolazione genica dovuta a fluttuazioni nei tempi di trascrizione e traduzione. Infine i modelli basati su equazioni differenziali [32–34] presentano il notevole vantaggio di considerare i dati in un range continuo di espressione e di poter rappresentare agevolmente i meccanismi di feedback positivo e negativo nel modello di regolazione.

I metodi basati sul confronto pair-wise dei geni si basano, invece, su un confronto dei profili di espressione di tutte le coppie di geni, alla ricerca di possibili relazioni di interazione messe in luce dai profili stessi. Si tratta di metodi relativamente più semplici e che, a differenza dei metodi basati su modello, non necessitano di un elevato numero di campioni per poter essere applicati. Esistono svariati metodi appartenenti a questa categoria, che si differenziano principalmente per la *misura di similarità* adottata: mediante tale misura viene valutato il grado di correlazione tra le coppie di profili di espressione genica [27]. Un primo esempio in letteratura è dato dalle *Relevance Networks* [35] in cui viene utilizzata come misura di similarità la mutua informazione tra i profili. Anche ARACNE [36] (*Algorithm for the Reconstruction of Accurate Cellular Networks*) calcola la mutua informazione di ogni coppia di geni ma estendendone la definizione a variabili continue, così da non dover discretizzare i dati. Altri metodi invece utilizzano misure di similarità basate sulla correlazione tra profili [37, 38] o sulla correlazione parziale [39–42]: si differenziano per dettagli implementativi quali il criterio per la selezione delle relazioni significative o eventuali operazioni di "pre-processing" dei dati (ad esempio l'applicazione di uno step di selezione dei geni differenzialmente espressi per eliminare i profili piatti o di clustering per raggruppare i profili di espressione identici). A differenza dei metodi basati su modello, i metodi basati su confronto pair-wise dei profili di espressione non possono essere considerati alla stregua di veri e propri metodi di reverse engineering per inferire sulla rete di regolazione genica. Sebbene misure di correlazione tra variabili vengano spesso utilizzate come indicatori di una relazione di tipo causa-effetto tra variabili altamente correlate, in realtà un elevato grado di correlazione tra profili di espressione non è necessariamente indicativo di una effettiva interazione di tipo regolatorio tra i

geni. Due geni possono essere ben correlati sia quando interagiscono in maniera diretta nei meccanismi di controllo e regolazione, ma anche quando interagiscono in maniera indiretta, o quando sono regolati da un regolatore comune, o, più semplicemente, quando sono contemporaneamente espressi all'interno della cellula, con profili di espressione simili, pur non essendo coinvolti nello stesso processo biologico. Tali metodi possono essere quindi pensati come metodi che consentono di individuare relazioni significative tra profili di espressione genica: tali relazioni sono certamente indicative del sottostante sistema di regolazione ma non necessariamente rappresentative di un'effettiva interazione tra i geni.

2.2 Confronto pair-wise tra profili di espressione

Nella prima fase del lavoro ci si è posti come obiettivo quello di individuare le relazioni significative tra geni a partire da dati di espressione statici e di costruire una rete che rappresenti le relazioni individuate. Si è scelto di adottare, per questa prima elaborazione dei dati di espressione genica, metodi basati su confronto pair-wise, utilizzando due diverse misure di similarità tra geni. Il motivo di tale scelta è legato, in primo luogo, al fatto che si tratta di metodi più semplici e che, a differenza dei metodi basati su modello, non fanno alcuna assunzione sui dati o sui modelli che descrivono i meccanismi di regolazione. Inoltre, mentre i metodi basati su modello richiedono in genere dati ottenuti mediante esperimenti ad hoc opportunamente disegnati e un numero di campioni talvolta molto elevato, i metodi basati su confronto pair-wise possono essere applicati, senza particolari restrizioni, a dati di espressione ottenuti in condizioni sperimentali diverse e non richiedono un numero elevato di campioni. Infine tali metodi sono particolarmente adatti allo scopo del lavoro proposto: l'obiettivo, infatti, non è tanto quello di ricostruire la vera rete di regolazione quanto piuttosto quello di individuare delle relazioni significative tra profili di espressione e analizzare come variano tali relazioni in condizioni diverse. In particolare si è interessati ad analizzare come varia il grado di correlazione tra geni confrontando una popolazione di soggetti sani con una popolazione di soggetti che presentano una qualche patologia, o due popolazioni di soggetti che presentano forme diverse di una stessa patologia.

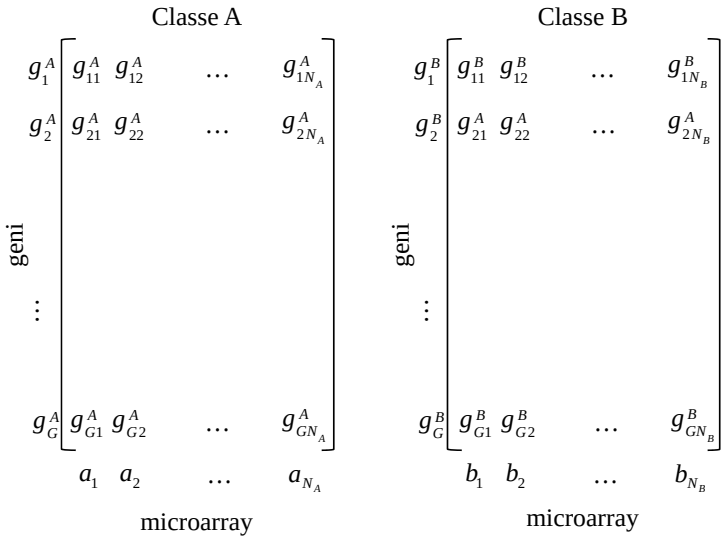
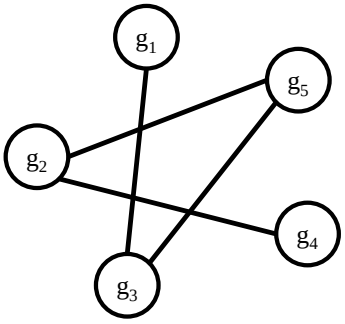


Figura 2.2 Rappresentazione delle matrici di valori di espressione genica per le due classi di soggetti: le righe corrispondono ciascuna ad uno specifico trascritto monitorato mentre le colonne agli array relativi ai diversi soggetti.

Per questo studio sono stati utilizzati dati di espressione genica di G geni, ottenuti per due distinte classi di popolazione A e B, caratterizzate rispettivamente da N_A e N_B soggetti, come schematizzato in figura 2.2. I dati utilizzati sono dati statici: ciascun campione dei profili fornisce una quantificazione del livello di mRNA nel tessuto, prelevato da uno dei soggetti monitorati, in una condizione di stato stazionario del sistema. Si è scelto di utilizzare dati di espressione statici piuttosto che dati dinamici, costituiti da profili di espressione temporali, poiché si tratta di dati più semplici da ottenere e che sono tipicamente disponibili con maggior facilità rispetto ai dati dinamici.

I dati di espressione utilizzati in questo lavoro sono dati simulati che consentono di procedere ad una validazione del metodo di analisi proposto: i dettagli sul simulatore adottato e sul procedimento di simulazione verrano presentati nel capitolo 4.

La rete di interazioni tra geni viene rappresentata mediante un grafo dove i vertici rappresentano i geni monitorati e gli archi le relazioni che intercorrono tra coppie di geni. Il grafo costruito è non orientato: infatti le relazioni tra profili di espressione sono simmetriche e prive di direzionalità, indipendentemente dalla misura di similarità adottata. Inoltre nel grafo non sono presenti *self-loop*, ovvero archi che congiungono un



	g_1	g_2	g_3	g_4	g_5
g_1	0	0	1	0	0
g_2	0	0	0	1	1
g_3	1	0	0	0	1
g_4	0	1	0	0	0
g_5	0	1	1	0	0

Figura 2.3 Esempio di rappresentazione grafica e matriciale di una rete di interazioni tra geni con $G = 5$.

nodo a se stesso: non sono infatti state considerate le relazioni dei profili di espressione con se stessi, che risulterebbero tutte caratterizzate dal valore di correlazione massimo e non sono significative per lo scopo di questo lavoro.

Le reti costruite vengono rappresentate in maniera compatta mediante la relativa matrice di adiacenza Q. Si tratta di una matrice quadrata di dimensione G e simmetrica, dato che il grafo è non orientato. Gli elementi di Q sono definiti come

$$q_{ij} = \begin{cases} 1 & \text{se i geni } g_i \text{ e } g_j \text{ sono congiunti da un arco} \\ 0 & \text{altrimenti.} \end{cases}$$

Un esempio esplicativo è riportato in figura 2.3. Nel seguito della trattazione si farà riferimento alla costruzione di una sola rete, sottintendendo che il procedimento di costruzione è identico per le due classi di soggetti.

2.3 Misure di correlazione

La rete di interazioni tra geni costruita varia notevolmente al variare della misura di similarità adottata per confrontare i profili di espressione: il lavoro di Soranzo *et al.* [27] presenta un confronto tra le misure più comunemente utilizzate.

In questo lavoro si è scelto di adottare due diverse misure, entrambe di tipo *covariance-based*. La *correlazione di Pearson* [43] è una misura ampiamente utilizzata per determinare il grado di relazione tra geni, tuttavia presenta alcune limitazioni

nell'individuazione dei pattern di interazione in un contesto multivariato. Si tratta infatti di una misura che può mettere in relazione anche geni che in realtà interagiscono indirettamente o attraverso l'azione di altri geni, definendo così un arco nel grafo che risulta essere un falso positivo rispetto alla rete reale.

Per far fronte a tale problema è possibile ricorrere alla nozione di indipendenza condizionata dalla teoria dei *Graphical Models* [44], calcolando la correlazione residua tra due variabili dopo aver condizionato rispetto alla conoscenza del profilo di espressione di uno o più altri geni. Per far ciò ci si avvale di una misura di confronto tra profili che prende il nome di *correlazione parziale di Pearson*.

Nonostante la misura di correlazione parziale consenta di individuare relazioni più affidabili di quelle individuate mediante correlazione [39], si è scelto di utilizzare entrambe le misure al fine di valutare se e quanto il metodo di analisi proposto per la selezione dei biomarcatori è sensibile al metodo adottato per individuare le relazioni tra geni.

2.3.1 Correlazione di Pearson

In molteplici lavori è stata utilizzata la correlazione di Pearson [43] per il confronto di profili di espressione genica [37, 38, 45]. Si tratta infatti di una misura che può essere applicata a dati continui e che richiede un numero relativamente poco elevato di campioni [46].

Definizione 2.3.1 *Date due variabili aleatorie x e y la loro* correlazione di Pearson r_{xy} *è data da*

$$r_{xy} := \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

dove σ_{xy} indica la covarianza tra le due variabili, mentre σ_x e σ_y indicano la deviazione standard rispettivamente di x e y.

Considerando ciascun gene g_i come una variabile aleatoria di cui si dispone di un numero limitato di campioni, $g_{in}, n = 1 \dots N$, la correlazione di Pearson tra i profili di espressione dei geni g_i e g_j può essere calcolata mediante il seguente stimatore:

$$r_{ij} = \frac{\frac{1}{N-1}\sum_{n=1}^{N}(g_{in} - \overline{g_i})(g_{jn} - \overline{g_j})}{\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(g_{in} - \overline{g_i})^2}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(g_{jn} - \overline{g_j})^2}}, \tag{2.1}$$

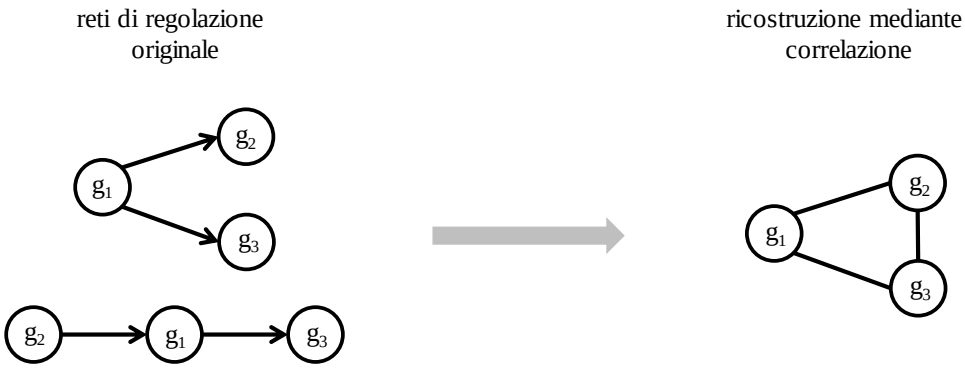


Figura 2.4 Errore di tipo falso positivo: i geni g_2 e g_3 presentano un elevato coefficiente di correlazione pur non interagendo in maniera diretta nella rete di regolazione originale.

dove N è il numero di campioni disponibili e $\overline{g_i}$ e $\overline{g_j}$ indicano la media campionaria dei profili di espressione di g_i e g_j.

La correlazione di Pearson individua relazioni di tipo lineare tra le variabili messe a confronto e assume valori compresi tra -1 e 1. Nel contesto di questo lavoro si prende in considerazione esclusivamente il valore assoluto della correlazione tra due variabili.

Tuttavia la correlazione di Pearson non è in grado di distinguere tra interazioni di tipo diretto e indiretto. Un elevato coefficiente di correlazione tra due variabili non necessariamente è indicativo di un'interazione diretta in cui uno dei due geni regola il livello di espressione dell'altro: due geni, infatti, possono essere altamente correlati anche nel caso in cui interagiscano in maniera indiretta, ad esempio per la presenza di un regolatore comune, o perché la loro interazione è mediata da una terza variabile in un pattern di interazione di tipo sequenziale. Queste due situazioni, rappresentate in figura 2.4, conducono alla selezione di relazioni nella rete costruita che non corrispondono ad una reale interazione tra geni, si commettono cioè degli errori di tipo falso positivo.

Nonostante questa limitazione intrinseca, la presenza di un elevato numero di falsi positivi, ai fini di questo lavoro, non preclude la possibilità di confrontare le relazioni geniche nelle reti costruite con l'obiettivo di analizzare le variazioni che intercorrono tra le due reti inferite dalle due classi di soggetti.

2.3.2 Correlazione parziale di Pearson

Per superare, almeno in parte, i limiti illustrati per la correlazione di Pearson, è possibile utilizzare come misura per il confronto tra geni un coefficiente di correlazione parziale [39, 42], che quantifica il grado di correlazione tra due variabili condizionato rispetto a una o più delle restanti variabili in gioco. Il numero di variabili rispetto al quale viene eseguito il condizionamento definisce l'ordine del coefficiente di correlazione parziale.

Definizione 2.3.2 *Date due variabili aleatorie x e y la loro* correlazione parziale di Pearson del primo ordine *è data da*

$$p_{xy}^{(l)} := \min_{z \neq x,y} |r_{xy|z}|,$$

dove

$$r_{xy|z} := \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

La correlazione parziale del primo ordine tra i geni g_i e g_j si ottiene cioè condizionando la coppia di variabili g_i, g_j rispetto a tutte le restanti variabili g_k, con $k \neq i, j$, prese singolarmente. Se esiste una variabile g_k capace di spiegare interamente la correlazione esistente tra le variabili g_i e g_j, la correlazione parziale tra g_i e g_j assume un valore prossimo a zero e le due variabili si dicono condizionatamente indipendenti dato g_k. Infatti la correlazione tra g_i e g_j condizionata rispetto a g_k è data dalla correlazione tra le parti di g_i e g_j che non sono correlate con g_k: se $r_{ij|k} \approx 0$ significa che la correlazione esistente tra g_i e g_j è unicamente dovuta alla correlazione che entrambi i geni presentano con la variabile g_k e i due geni possono essere considerati condizionatamente indipendenti.

Utilizzando questo nuovo criterio per la costruzione della rete, che consiste nella selezione delle relazioni tra le coppie di geni che presentano un coefficiente di correlazione parziale significativamente diverso da zero, è dunque possibile distinguere le interazioni geniche di tipo diretto da quelle di tipo indiretto. Si pensi ad esempio alla situazione rappresentata in figura 2.5 in cui i due geni g_2 e g_3 presentano il regolatore comune g_1. Utilizzando come misura di confronto tra profili la correlazione di Pearson, tali geni, sebbene interagiscano solo indirettamente, vengono messi in relazione nella

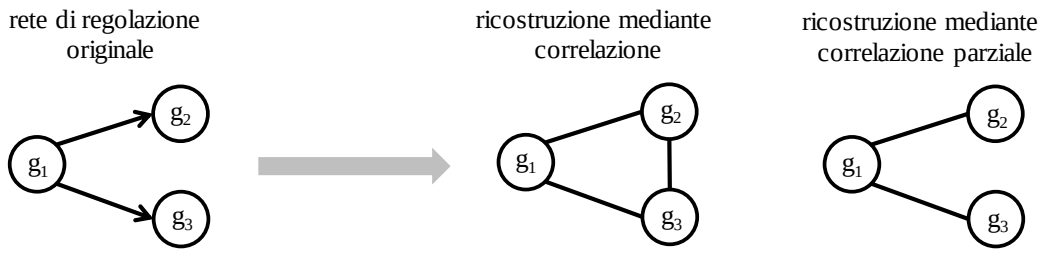


Figura 2.5 La misura di correlazione parziale, a differenza della correlazione semplice, è in grado di distinguere le relazioni tra geni di tipo diretto (tra g_1 e g_2 e tra g_1 e g_3) da quelle di tipo indiretto (tra g_2 e g_3).

rete ricostruita a causa del loro elevato coefficiente di correlazione. La correlazione parziale tra g_2 e g_3 assume invece un valore prossimo a zero evidenziando il fatto che l'elevata correlazione tra g_2 e g_3 è in realtà legata alla presenza di un regolatore comune: utilizzando questa misura per la ricostruzione della rete i geni g_2 e g_3 non vengono messi in relazione, mentre rimangono correttamente individuate le relazioni tra g_1 e g_2 e tra g_1 e g_3 caratterizzati da un'azione regolatoria diretta.

La definizione di correlazione parziale può essere estesa per calcolare la correlazione parziale di ordini superiori, qualora il condizionamento rispetto ad una sola variabile non fosse sufficiente. Ad esempio la correlazione parziale del secondo ordine tra due variabili g_i e g_j si ottiene condizionando rispetto a tutte le possibili coppie di variabili g_k e g_l, con $k, l \neq i, j$, ed è definita da:

$$p_{ij}^{(2)} = \min_{k,l \neq i,j} |r_{ij|kl}|,$$

dove

$$r_{ij|kl} = \frac{r_{ij|k} - r_{il|k} r_{jl|k}}{\sqrt{(1 - r_{il|k}^2)(1 - r_{jl|k}^2)}}.$$

In questo modo è possibile escludere dalla rete ricostruita relazioni tra variabili che possono essere spiegate dalla presenza di una coppia di regolatori comuni. Alcuni lavori hanno dimostrato che, nell'analisi di dati di espressione genica, è sufficiente considerare misure di correlazione parziale fino al secondo ordine per individuare relazioni affidabili [39]. Tuttavia in questo contesto si è scelto di applicare un condizionamento

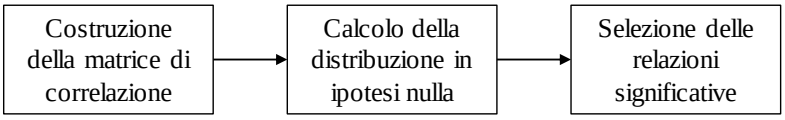


Figura 2.6 Fasi principali del procedimento di costruzione della rete di interazioni geniche.

esaustivo, condizionando la correlazione tra due geni rispetto a tutte le restanti $G - 2$ variabili. Nel resto della trattazione si indicherà con p_{ij} la correlazione parziale tra i profili g_i e g_j condizionata rispetto a tutte le restanti variabili.

2.4 Costruzione della rete di interazioni geniche

Per procedere alla costruzione della rete delle interazioni tra geni, una volta scelte le misure di similarità da adottare per il confronto dei profili di espressione, è necessario introdurre un criterio che consenta di selezionare solo le relazioni geniche significative, caratterizzate da un alto grado di correlazione (o correlazione parziale). Nelle prossime sezioni verrà illustrato nel dettaglio il procedimento di costruzione della rete, riassunto nello schema riportato in figura 2.6.

Il codice per la costruzione delle reti di interazioni tra geni è stato scritto con linguaggio di programmazione R [47]. In particolare sono state utilizzate alcune funzioni, per il calcolo della matrice di correlazione parziale e delle distribuzioni in ipotesi nulla, implementate nel pacchetto GeneNet (versione 1.2.4), scaricabile dall'archivio di Bioconductor [48].

2.4.1 Costruzione della matrice di correlazione e correlazione parziale

I valori di correlazione calcolati per tutte le coppie di geni vengono memorizzati in una matrice di correlazione R il cui elemento in posizione $[i, j]$ corrisponde alla correlazione r_{ij} tra la coppia di geni g_i e g_j. Le correlazioni di Pearson tra tutte le coppie di geni possono essere calcolate in modo immediato applicando la definizione (2.1). La matrice ottenuta è una matrice simmetrica poiché $r_{ij} = r_{ji}, \forall i, j = 1, 2, \ldots, G$, e gli elementi

sulla diagonale, ovvero i valori di correlazione di ciascun gene con se stesso, sono tutti unitari.

Il calcolo delle correlazioni parziali è invece più laborioso, poiché richiede il calcolo della correlazione tra le coppie di geni condizionata alla conoscenza di tutti i profili rimanenti. La correlazione parziale tra i due profili g_1 e g_2 può essere calcolata come la correlazione tra i residui ϵ_1 e ϵ_2 che risultano dalla regressione lineare dei geni g_1 e g_2 su tutti i profili rimanenti, $g_3, g_4, \ldots, g_G$. Assumendo che i dati di espressione siano estratti da una distribuzione normale multivariata e che la matrice di correlazione R sia invertibile, si può dimostrare che la matrice di correlazione parziale P è legata all'inversa di R [44]. Sia $\Omega = R^{-1}$ l'inversa della matrice di correlazione, gli elementi di P possono essere calcolati mediante la seguente relazione:

$$p_{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, \tag{2.2}$$

dove ω_{ij} è l'elemento in posizione $[i, j]$ della matrice Ω.

Tuttavia quando il numero di variabili analizzate G è maggiore del numero di campioni disponibili N ($N \ll G$), situazione molto frequente per i dati di espressione ricavati mediante microarray, la matrice R non è definita positiva, quindi non è invertibile e la (2.2) non può essere applicata. In questi casi è possibile utilizzare per il calcolo della matrice P lo stimatore proposto da Schäfer e Strimmer in [42], appositamente ideato per questo tipo di problema. Tale stimatore fa uso della matrice pseudoinversa di Moore-Penrose per l'inversione della matrice di correlazione R e si serve delle tecniche di *bootstrap* e *bagging* per stabilizzare la stima ottenuta.

La matrice pseudoinversa di Moore-Penrose [49] è una generalizzazione della matrice inversa che può essere applicata anche a matrici singolari e che si basa sulla decomposizione ai valori singolari (SVD). La matrice di correlazione R può essere fattorizzata come $R = UDV^T$, dove D è una matrice quadrata diagonale i cui elementi singolari sono tutti strettamente positivi. La pseudoinversa R^+ è definita come $R^+ = VD^{-1}U^T$ e richiede soltanto l'inversione della matrice diagonale D. Si può dimostrare che la pseudoinversa R^+ è la soluzione ai minimi quadrati dell'equazione $RR^+ = I$ e che, quando R è invertibile, coincide con la matrice inversa R^{-1}.

La tecnica del bagging è una tecnica che viene frequentemente utilizzata per stabilizzare il valore di una stima e ridurne la varianza. Si consideri un generico stimatore $\hat{\theta}(y)$ applicato ad un generico dataset y, la tecnica del bagging può essere così schematizzata:

1. Si applica un approccio bootstrap per generare B nuovi dataset $y^b, b = 1, \ldots, B$, della stessa dimensione di y, campionando con ripetizione dal dataset originale;
2. Per ogni dataset y^b così ottenuto si calcola la stima $\hat{\theta}^b$;
3. La stima finale si ottiene mediando le B stime calcolate: $\hat{\theta} = (1/B) \sum_{b=1}^{B} \hat{\theta}^b$.

Lo stimatore per la matrice di correlazione parziale $\hat{P}$ presentato in [42] utilizza il bagging per stimare la matrice di correlazione R a partire dai dati e calcola poi la pseudoinversa della matrice così ottenuta per ottenere una stima di P.

2.4.2 Calcolo della distribuzione in ipotesi nulla

L'applicazione dei test statistici per sottoporre a verifica l'ipotesi $H_0 : r_{ij} = 0$ e $H_0 : p_{ij} = 0$ per ogni coppia di geni, richiede la conoscenza delle distribuzioni dei coefficienti di correlazione e correlazione parziale in ipotesi nulla, a partire dalle quali è possibile calcolare i p-value associati a ciascuna coppia di geni. La distribuzione in ipotesi nulla dei coefficienti di correlazione semplice, r, è nota in forma esatta [50] ed è data da:

$$f_0(r; \kappa) = (1 - r^2)^{(\kappa-3)/2} \frac{\Gamma(\kappa/2)}{\pi^{1/2}\Gamma[(\kappa - 1)/2]}, \tag{2.3}$$

dove Γ è la funzione gamma di Eulero [51]. L'unico parametro da cui dipende tale distribuzione è κ, ovvero il numero di gradi di libertà, che è legato al numero di campioni N disponibili per ciascuna variabile dalla relazione $\kappa = N - 1$. Un'esempio della forma che tale distribuzione assume per $\kappa = 11$ è riportato in figura 2.7: come è ovvio attendersi i coefficienti di correlazione assumono con probabilità non nulla valori compresi tra -1 e 1 e i valori assunti con maggior probabilità sono quelli prossimi a zero.

Si può dimostrare che i coefficienti di correlazione parziale sono distribuiti esattamente come i coefficienti di correlazione semplice, l'unica differenza è data dal numero

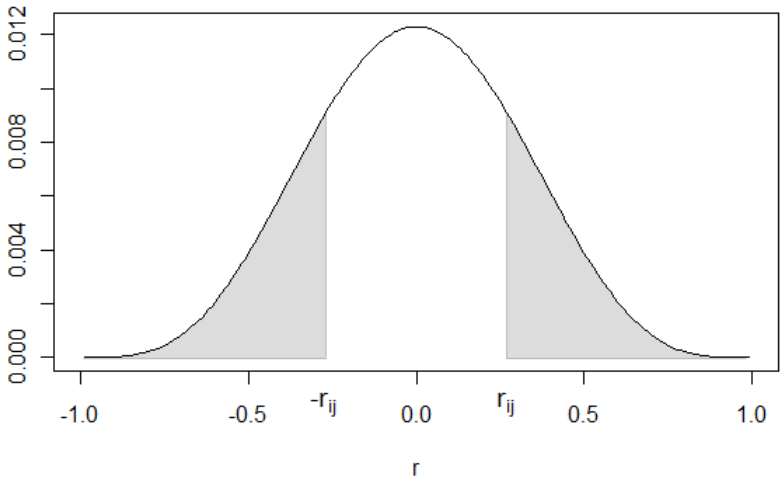


Figura 2.7 Esempio di distribuzione dei coefficienti di correlazione r in ipotesi nulla per $N = 12$ e $\kappa = 11$: l'area grigia corrisponde al p-value associato alla coppia di geni con correlazione r_{ij}.

di gradi di libertà κ che deve essere ridotto del numero di variabili rispetto alle quali si esegue il condizionamento nel calcolo della correlazione parziale [50]. Se G è il numero di variabili considerate, eseguendo un condizionamento esaustivo rispetto a $G-2$ variabili il numero di gradi di libertà è dato da $\kappa = N - G + 1$. Questa relazione è valida solamente se il numero di variabili G non supera il numero di campioni disponibili N, altrimenti κ assume un valore negativo. Nel caso in cui $N < G$ è necessario ricorrere allo stimatore $\hat{P}$ per il calcolo dei coefficienti di correlazione parziale tra le variabili: si può dimostrare che anche in questo caso la distribuzione dei coefficienti in ipotesi nulla è definita dall'equazione (2.3), ma con $\kappa \geq 0$ anche per $N < G$. Il numero di gradi di libertà κ non è più una semplice funzione di N e G ma viene stimato a partire dai dati mediante un procedimento non banale descritto in [42].

2.4.3 Selezione delle relazioni significative

Una volta costruita la matrice di correlazione e calcolata la distribuzione in ipotesi nulla, si ricorre all'applicazione di test statistici per valutare la significatività dei valori di correlazione calcolati. Se G è il numero di geni ed $E = G(G-1)/2$ il numero di tutte le possibili coppie di geni e dei potenziali archi che li congiungono nella rete, sarà

necessario applicare E test statistici per determinare la presenza o meno degli archi nella rete sulla base dei relativi valori di correlazione. L'ipotesi nulla $H_0 : r_{ij} = 0$ viene sottoposta a verifica per ciascuna coppia di geni: se H_0 viene rifiutata si può dire che il livello di correlazione tra i geni è significativamente diverso da zero e g_i e g_j vengono congiunti con un arco, se invece l'ipotesi nulla viene accettata i geni sono considerati non correlati e non viene introdotto alcun arco nella rete.

A partire dalla distribuzione in ipotesi nulla e dai valori di correlazione calcolati per i dati, è possibile calcolare per ciascuna coppia di geni il relativo p-value, che rappresenta la probabilità di rifiutare l'ipotesi nulla quando in realtà questa è verificata. Il p-value associato ad una coppia di geni corrisponde cioè alla probabilità di commettere un errore di tipo falso positivo selezionando la relazione tra i geni con un arco nella rete, quando in realtà i geni non sono correlati. Per selezionare le relazioni significative è dunque necessario definire qual è la probabilità di commettere un errore di tipo falso positivo che si è disposti ad accettare, che prende il nome di livello di significatività α. Se il p-value relativo alla coppia g_i, g_j è inferiore ad α l'ipotesi nulla viene rifiutata e la relazione tra i due geni viene selezionata nella rete. In alternativa è possibile fissare la soglia di confidenza θ corrispondente al livello di significatività α in modo tale che:

- se $|r_{ij}| \geq \theta$, la correlazione tra i geni g_i e g_j viene considerata significativa e i due geni vengono uniti da un arco nella rete;
- se $|r_{ij}| < \theta$, la correlazione tra g_i e g_j non è significativa e non viene inserito alcun arco nella rete.

Nella scelta del livello di significatività α è opportuno tenere in considerazione il fatto che limitando eccessivamente la soglia sulla probabilità, detta FP-rate, di commettere errori di tipo falso positivo che si è disposti ad accettare, si riscontra necessariamente un incremento della probabilità di commettere errori di tipo falso negativo: se α è troppo piccolo aumentano le coppie di geni che, pur essendo realmente correlate, non vengono selezionate perché caratterizzate da p-value maggiore di α.

Il criterio di selezione descritto fin'ora, basato sul controllo della FP-rate del singolo test statistico, non è adeguato per test statistici multipli, come in questo caso dove viene applicato un test statistico per ciascuno degli E potenziali archi della rete da

costruire. Risulta quindi opportuno applicare appositi criteri per la determinazione della soglia di confidenza, definiti correzioni per test multipli, che consentono di controllare il valore della FP-rate globale, ovvero la probabilità di commettere un errore di tipo falso positivo sugli E test piuttosto che sul singolo test. In questo lavoro si è adottata la correzione per test multipli basata su *false discovery rate* [52], i cui dettagli sono presentati nell'appendice A.

3

Metodi per il ranking dei biomarcatori

L'identificazione dei biomarcatori a partire da dati di espressione genica è un problema molto studiato nel campo della bioinformatica per l'interesse biologico e clinico legato alla conoscenza dei geni in grado di caratterizzare una particolare patologia. In questo capitolo viene presentato un nuovo metodo per l'identificazione robusta di biomarcatori a partire da dati di espressione genica. Nella prima parte viene introdotto il problema dell'identificazione dei biomarcatori presentando le metodologie più comunemente utilizzate. Nella seconda parte, invece, viene introdotto il nuovo metodo di ranking proposto, ponendo in particolare l'attenzione sulle caratteristiche che lo distinguono dalle tecniche attualmente utilizzate.

3.1 Definizione di biomarcatore

Negli ultimi 15-20 anni la diffusione delle tecnologie high-throughput quali i microarray per l'analisi del trascrittoma ha rivoluzionato l'approccio allo studio di diverse malattie di origine genetica, soprattutto di malattie genetiche complesse come ad esempio i tumori. Tipicamente, in un disegno sperimentale, i dati di espressione da analizzare provengono da soggetti diversi e/o con fenotipi diversi. Il confronto e l'analisi dei dati così ottenuti consentono di indagare i meccanismi che caratterizzano il manifestarsi e l'evolvere di diverse patologie. Per alcune malattie (come ad esempio il cancro o le malattie neurodegenerative) l'analisi di questi dati si è dimostrata estremamente utile per l'identificazione di geni, detti biomarcatori, in grado di determinare lo stato patologico di un soggetto malato e per lo sviluppo di nuove ipotesi sulla fisiologia utili per rispondere a domande di tipo diagnostico, prognostico e funzionale.

Di particolare interesse in campo medico e biologico è l'identificazione dei biomarcatori di patologie complesse, come i tumori, caratterizzate dalla presenza di numerose mutazioni e alterazioni funzionali a carico di diversi geni, piuttosto che da alterazioni di

un singolo gene o cromosoma tipiche della gran parte delle malattie ereditarie. Questa intrinseca complessità della malattia rende particolarmente utile ed interessante poter disporre di una lista di biomarcatori che la caratterizzano e ne descrivono l'evoluzione. D'altro canto questa stessa complessità presente nei meccanismi biologici che descrivono tali patologie e la presenza di un certo grado di eterogeneità che contraddistingue le diverse classi di malattia fanno si che la lista di biomarcatori sia difficile da definire.

Sebbene una vera e propria definizione di biomarcatore non esista, è restrittivo considerare come biomarcatori esclusivamente dei geni o delle porzioni di genoma affette da mutazioni o alterazioni. Possono essere infatti considerati biomarcatori non solo porzioni di DNA, ma anche molecole di altra natura, come molecole di mRNA, proteine e metaboliti, o processi biologici, quali apoptosi cellulare, angiogenesi e proliferazione incontrollata [4, 53]. In realtà, procedendo alla ricerca dei biomarcatori a partire da dati di espressione genica, non si individuano direttamente le molecole e i processi biologici che caratterizzano la malattia quanto piuttosto i geni che li codificano. Per questo motivo si fa spesso riferimento a geni biomarcatori. Dall'analisi di dati ottenuti mediante microarray è possibile mettere a confronto i profili di espressione di due popolazioni di soggetti caratterizzati da un particolare stato fisiologico/patologico permettendo di distinguere anche sottoclassi della stessa malattia. In questo modo si individuano i geni la cui attività è stata alterata e che risultano sovraregolati o sottoregolati nei soggetti malati rispetto a quelli di controllo.

La classificazione di un soggetto, in base al suo profilo di espressione, come appartenente ad una precisa classe che caratterizza lo stato clinico consente di somministrare una terapia specifica per la patologia analizzata, massimizzandone l'efficacia e minimizzandone la tossicità [54]. I biomarcatori sono di fondamentale importanza anche in fase diagnostica, poiché un loro frequente monitoraggio consente di effettuare una diagnosi precoce dell'eventuale presenza di una patologia. In questo modo è possibile somministrare la terapia quando la malattia non è ancora ad uno stadio avanzato ed è più facilmente guaribile. Una volta diagnosticata la malattia, i biomarcatori possono essere utilizzati per valutare fattori clinici legati alla sua evoluzione. In fase prognostica, invece, i biomarcatori possono essere utilizzati per prevedere l'andamento della malattia

o per predire la risposta dei pazienti ai trattamenti somministrati in modo tale da poter modificare e migliorare gli schemi terapeutici previsti. I biomarcatori possono essere inoltre impiegati per valutare l'efficacia di un farmaco e i suoi meccanismi d'azione: osservando come varia l'espressione genica in risposta alla somministrazione di un farmaco è possibile stabilirne l'indice terapeutico e individuare eventuali effetti tossici. Nel prossimo futuro la conoscenza dei biomarcatori genici porterà allo sviluppo di terapie mirate basate sull'utilizzo di farmaci in grado di agire in modo specifico sulle molecole alterate dalla malattia. La possibilità di individuare in maniera precisa e specifica i biomarcatori genici sarà dunque cruciale per identificare un bersaglio terapeutico appropriato su cui intervenire [3].

3.2 Le tecniche di feature selection

Le molteplici applicazioni e l'interesse biologico, medico e clinico legati alla possibilità di conoscere i biomarcatori di patologie complesse come i tumori hanno fatto si che negli ultimi anni siano state proposte numerosissime tecniche per la ricerca e la selezione di biomarcatori a partire da dati di espressione genica. Tali tecniche, dette di *feature selection*, consentono di individuare, tra le decine di migliaia di variabili monitorate, le variabili significative del processo biologico o della patologia in fase di studio. Nella maggior parte dei casi si tratta di tecniche di feature selection proposte e sviluppate in altri ambiti, come ad esempio nel campo di *machine learning* e *data mining*, e poi modificate e riadattate per poter essere applicate a dati di origine biologica.

In un classico problema di feature selection l'obiettivo è quello di individuare, a partire dai dati di espressione di due classi di soggetti malati e di controllo o appartenenti a classi di malattia diverse, le variabili che risultano significativamente differenti nei due gruppi o che consentono di discriminare meglio tra le due classi di individui. Tipicamente il numero di variabili (geni) monitorate è di gran lunga superiore al numero di campioni disponibili per ciascuna variabile, che corrisponde al numero di soggetti di ciascuna classe.

Gli algoritmi di feature selection possono essere suddivisi in due categorie principali: la categoria dei *filter methods* e quella dei *wrapper methods* [55]. I filter methods

determinano l'importanza di ciascuna variabile tenendo conto esclusivamente del contenuto informativo e delle proprietà intrinseche dei dati. I wrapper methods, invece, selezionano, tra le variabili iniziali, un sottoinsieme di variabili significative risolvendo un problema di classificazione.

Le tecniche di feature selection appartenenti alla prima categoria prevedono l'assegnazione di un punteggio proporzionale, secondo un prefissato criterio, alla rilevanza di ciascuna variabile: solo le variabili con punteggio al di sopra di una certa soglia vengono considerate significative. Le variabili così selezionate possono essere utilizzate, in un secondo momento, come variabili di ingresso ad un classificatore per suddividere i soggetti nelle due classi di appartenenza, ma la selezione delle variabili resta un procedimento del tutto indipendente dall'ottimizzazione delle prestazioni del classificatore. I filter methods, applicati alla selezione di biomarcatori a partire da dati di espressione genica, identificano come biomarcatori i geni che risultano essere differenzialmente espressi nelle due classi di soggetti. Per far ciò è necessario applicare un test statistico che consenta di ordinare i geni in base a quanto significativamente diverso è il loro livello di espressione nelle due classi di soggetti osservati. Le tecniche di feature selection basate sulla selezione dei geni differenzialmente espressi possono essere ulteriormente suddivise in due sottoclassi: i metodi parametrici e i metodi non parametrici. I metodi parametrici fanno delle precise assunzioni circa la distribuzione dalla quale sono stati campionati i dati di espressione osservati. Tra i test statistici parametrici più frequentemente utilizzati nell'analisi di dati di microarray vi sono il *t*-test e il test ANOVA [12]. In alternativa, quando non sono disponibili informazioni circa la distribuzione dei dati o quando è difficile verificare le assunzioni a causa della scarsità di campioni disponibili, è possibile ricorrere all'applicazione di metodi non parametrici, o *model-free*. Questi metodi non fanno alcuna ipotesi, o comunque fanno ipotesi meno stringenti, circa la distribuzione statistica dei dati. Il test dei ranghi di Wilkoxon [56] e il metodo del prodotto dei ranghi [57] ne sono un esempio. Una particolare classe di metodi non parametrici consente di stimare la statistica di riferimento del test a partire da permutazioni random ripetute e indipendenti dei dati stessi (ad esempio il test SAM [13]). I filter methods possono essere sia univariati che multivariati, i primi

considerano ciascuna variabile singolarmente, ignorando completamente l'esistenza di possibili relazioni tra le variabili geniche, i secondi invece prendono in considerazione coppie [14] o sottoinsiemi [15] di variabili.

Le tecniche di feature selection appartenenti alla classe dei wrapper methods, a differenza delle metodologie descritte precedentemente, selezionano iterativamente le variabili che meglio discriminano i soggetti delle due classi e che massimizzano le prestazioni di un classificatore nel classificare correttamente i soggetti. Metodi di questo tipo prendono in considerazione dei sottoinsiemi di variabili e la loro capacità di discriminare i soggetti appartenenti alle due classi ricorrendo all'utilizzo di test statistici multivariati. Durante la procedura di selezione delle variabili e di ottimizzazione delle prestazioni del classificatore, a ciascuna variabile viene assegnato un peso indicativo della sua capacità di discriminare le due classi di soggetti. In questo modo, al termine dell'algoritmo di ottimizzazione, le variabili possono essere ordinate in base al peso ad esse associato e dunque alla loro significatività in un contesto multivariato. Le diverse tecniche appartenenti alla classe dei wrapper methods si distinguono principalmente in base al modello di classificazione adottato. Una tecnica molto spesso utilizzata nell'analisi di dati di microarray consiste nell'applicazione delle *Support Vector Machine* per la classificazione dei soggetti in base ai loro valori di espressione genica [58, 59]. Probabilmente a causa dell'elevata complessità computazionale dei wrapper methods, i filter methods sono, al giorno d'oggi, i metodi di feature selection più frequentemente utilizzati nel campo della bioinformatica.

3.3 Integrazione dell'informazione relativa alle interazioni geniche nelle tecniche di feature selection

I metodi ad oggi più diffusi per la selezione dei biomarcatori individuano come variabili significative del processo patologico analizzato i geni che discriminano meglio le diverse classi di soggetti. Il fatto di non tenere in nessun modo in considerazione le relazioni di interazione tra le variabili è sicuramente una delle limitazioni principali di tali metodi. I metodi multivariati sono in grado di tenere conto di come le variabili contribuiscono tra

loro per spiegare i dati, ma senza considerare esplicitamente le vere e proprie relazioni di interazione tra geni. Potrebbe quindi essere interessante introdurre, nel processo di selezione dei biomarcatori, informazioni circa le interazioni tra le variabili geniche, le quali hanno un ruolo cruciale nei meccanismi di regolazione e controllo dell'espressione genica.

In questo lavoro è stato sviluppato un nuovo metodo per la selezione dei biomarcatori che integri le informazioni sulle interazioni tra variabili geniche. Un generico processo patologico, infatti, può alterare il livello di espressione dei geni colpiti dalla malattia che si ripercuote sulla rete di regolazione genica. É dunque ragionevole attendersi che i geni biomarcatori siano caratterizzati non soltanto da un livello di espressione significativamente diverso nelle due classi di soggetti considerate, ma anche da un diverso grado e una diversa modalità di interazione con le altre variabili in gioco.

Il nuovo metodo proposto per la selezione dei biomarcatori può essere considerato nella categoria dei filter methods poiché assegna a ciascuna variabile un punteggio, indicativo della sua significatività, esclusivamente sulla base del contenuto informativo dei dati. La tecnica proposta individua i geni biomarcatori basandosi sul confronto dei dati di espressione genica di due classi di soggetti ma, a differenza delle tecniche standard, non seleziona i geni differenzialmente espressi quanto piuttosto i geni che esibiscono una differenza significativa di interazione con gli altri geni in due classi di soggetti A e B.

Ad ogni gene viene assegnato un punteggio, o *score*, basato sulle variazioni tra le interazioni nelle reti di interazioni geniche per le due classi di soggetti, costruite sulla base dei dati di espressione genica come descritto nel capitolo 2. In particolare i nodi di ciascuna rete sono associati in maniera biunivoca ai G geni monitorati, mentre gli archi indicano la presenza di una relazione di interazione significativa tra i due geni che uniscono. Per quantificare le variazioni delle interazioni del gene g_i con le restanti variabili si ricorre ad un confronto della topologia locale delle due reti, A e B, in un intorno del nodo associato al gene g_i. Le misure adottate per confrontare la topologia locale delle reti, descritte in seguito, consentono di individuare i nodi della rete che hanno modificato in maniera sensibile le loro connessioni con i restanti

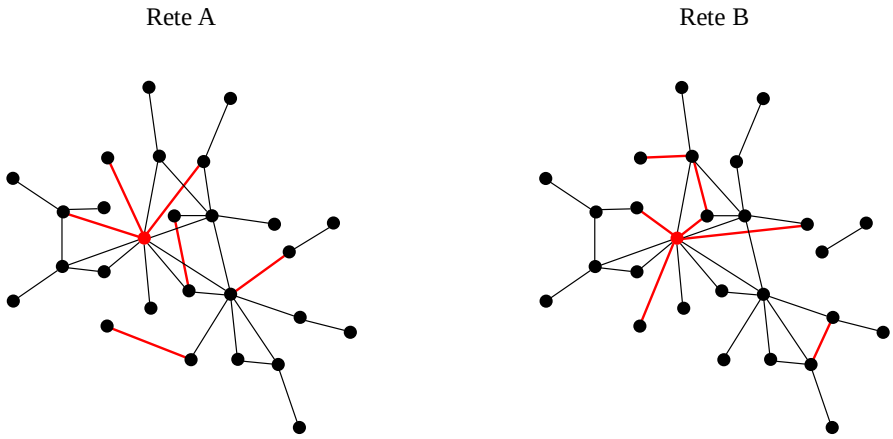


Figura 3.1 Rappresentazione grafica delle variazioni di interazioni tra geni. Gli archi in rosso rappresentano le relazioni di interazione tra geni che differiscono nelle due classi.

nodi. Se le misure adottate per il confronto della topologia locale sono sufficientemente robuste, ci si attende che i nodi così individuati siano associati a geni biomarcatori la cui attività e il cui grado di interazione con gli altri geni sono stati significativamente alterati dall'evolvere della malattia. Una rappresentazione grafica di ciò che ci si attende di osservare nelle reti di interazioni geniche costruite è riportata in figura 3.1: il gene associato al nodo indicato in rosso viene considerato un biomarcatore poiché caratterizzato da una consistente variazione di interazioni con gli altri geni.

3.4 Misura delle variazioni di interazione tra geni

Per confrontare le relazioni di interazione che ciascun gene presenta con le altre variabili geniche nelle due classi di soggetti e nelle relative reti di interazioni geniche costruite e quantificarne le variazioni, occorre definire una misura adeguata. Tale misura deve essere robusta sia al rumore di cui sono tipicamente affetti i dati di espressione ottenuti mediante microarray sia alla presenza dell'elevato numero di relazioni di tipo falso positivo individuate durante il procedimento di costruzione delle reti. É necessario, dunque, disporre di una misura che sia in grado di rilevare le principali variazioni di interazioni e che ridimensioni, rendendole trascurabili, le piccole variazioni legate alla presenza di rumore nei dati di espressione. Una misura con queste caratteristiche

assegna, idealmente, un punteggio elevato ai geni biomarcatori, i quali hanno subito una variazione significativa delle interazioni con gli altri geni a causa del manifestarsi dei meccanismi patologici.

In questo lavoro sono state proposte e messe a confronto quattro diverse misure per la quantificazione delle variazioni delle interazioni geniche: le prime tre sono basate su un confronto della topologia locale delle reti di interazioni geniche costruite, l'ultima, invece, si basa sulla valutazione della variazione del valore di correlazione, ovvero del peso attribuito ad ogni relazione tra coppie di geni.

3.4.1 Grado di un nodo

Una semplice quantificazione delle variazioni topologiche di un gene con gli altri geni presenti nella rete si ottiene confrontando il grado dei nodi associati al gene nelle due reti di interazioni. Poiché le reti di interazioni geniche costruite sono reti non orientate e non vi è alcuna distinzione tra archi entranti o archi uscenti, il grado di un nodo è definito come il numero di connessioni che il nodo presenta con altri nodi della rete.

Definizione 3.4.1 *Sia g_i l'i-esima delle G variabili geniche considerate e siano DEG_i^A e DEG_i^B rispettivamente il grado del nodo corrispondente a g_i nelle reti di interazioni geniche A e B. Lo* score basato sul grado *associato al gene g_i è definito da:*

$$S_{i,degree} = \frac{|log(DEG_i^A + 1) - log(DEG_i^B + 1)|}{log(max\{DEG_i^A, DEG_i^B\} + 2)}. \tag{3.1}$$

Per uniformità con la definizione dello score basato su graphlet, che verrà introdotto nella sezione 3.4.3, è stata applicata la trasformazione logaritmica e la normalizzazione, in modo tale da ottenere uno score che assume valori nel range $[0, 1)$.

Lo score così definito, basato sul confronto del grado dei nodi associati ad uno stesso gene, quantifica le variazioni di interazioni geniche tenendo in considerazione unicamente la variazione del numero di interazioni che ciascun gene presenta con gli altri geni nelle due reti identificate.

3.4.2 Coefficiente di clustering

Un'altra alternativa per la quantificazione delle variazioni di interazioni geniche è quella di ricorrere al confronto del *coefficiente di clustering* dei nodi. Il coefficiente di clustering è un parametro che descrive la tendenza dei nodi connessi ad uno stesso nodo a formare connessioni tra loro.

Definizione 3.4.2 *Sia g_i uno dei geni monitorati, il* coefficiente di clustering *del nodo associato a g_i in una delle reti costruite è definito da:*

$$CC_i = \frac{2q_i}{DEG_i(DEG_i - 1)},$$

dove DEG_i è il grado di connettività del nodo associato a g_i e q_i è il numero di connessioni presenti tra i nodi ad esso direttamente connessi.

Il coefficiente di clustering di un nodo è rappresentativo del rapporto tra il numero di interazioni tra i nodi ad esso direttamente connessi e il numero totale di possibili interazioni che tali nodi potrebbero formare tra loro. Si osservi, come esempio esplicativo, la figura 3.2.

Lo score associato alla variabile genica g_i può dunque essere definito come la differenza tra i coefficienti di clustering dei nodi associati a g_i nelle due reti di interazioni geniche.

Definizione 3.4.3 *Sia g_i l'i-esima delle G variabili geniche considerate e siano CC_i^A e CC_i^B rispettivamente il coefficiente di clustering del nodo corrispondente a g_i nelle reti di interazioni geniche A e B. Lo* score basato sul coefficiente di clustering *associato al gene g_i è definito da:*

$$S_{i,cc} = \frac{|log(cc_i^A + 1) - log(cc_i^B + 1)|}{log(max\{cc_i^A, cc_i^B\} + 2)}. \tag{3.2}$$

Anche in questo caso lo score assume valori nell'intervallo $[0, 1]$.

Lo score basato sul coefficiente di clustering, a differenza dello score basato sul grado dei nodi, non tiene conto della variazione delle interazioni dirette di ciascun gene con gli altri geni, quanto piuttosto della variazione delle relazioni che intercorrono tra i geni ad esso connessi. Tale misura può essere ragionevolmente utilizzata per la

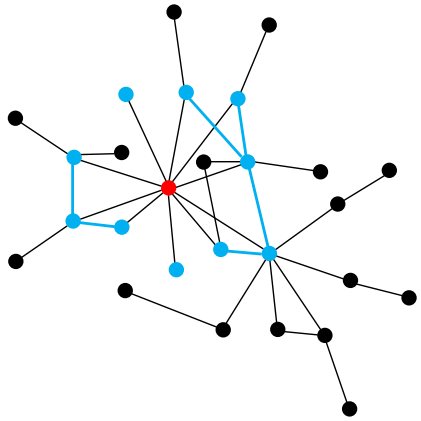


Figura 3.2 Rappresentazione grafica del coefficiente di clustering di un nodo. I 10 nodi direttamente connessi al nodo rosso, rappresentati in azzurro, formano tra loro 6 connessioni (archi azzurri), mentre il numero totale di connessioni che potrebbero formare è dato da $10 * 9/2$. Il coefficiente di clustering del nodo rosso è dunque dato da $CC = 6 * 2/10 * 9 = 0,1\bar{3}$.

selezione dei geni biomarcatori: é infatti sensato attendersi che i biomarcatori, oltre a modificare le loro interazioni con gli altri geni, diano luogo anche ad una variazione significativa delle relazioni esistenti tra i geni da essi direttamente regolati.

3.4.3 Graphlet

Per la definizione di una misura di confronto della topologia locale più complessa si è fatto riferimento alle nozioni di *graphlet* e di *signature* proposte da Pržulj e Milenković in [22, 60] nell'ambito dell'analisi delle reti di interazione proteica, o reti PPI (*Protein-Protein Interaction Networks*). La nozione di graphlet è stata introdotta, in questi lavori, come strumento per confrontare la struttura di reti biologiche in termini di proprietà topologiche locali. L'applicazione di queste misure a reti PPI reali ha dato risultati significativi dal punto di vista biologico che hanno dimostrato l'efficacia degli strumenti utilizzati per l'analisi strutturale delle reti. La ricerca di nodi con proprietà topologiche locali simili, sulla base della nozione di graphlet, ha infatti portato all'individuazione di insiemi di nodi associati a proteine con funzioni simili o coinvolte negli stessi processi biologici. I risultati ottenuti hanno dunque confermato la stretta relazione esistente tra le proprietà strutturali e topologiche delle reti biologiche e le caratteristiche funzionali

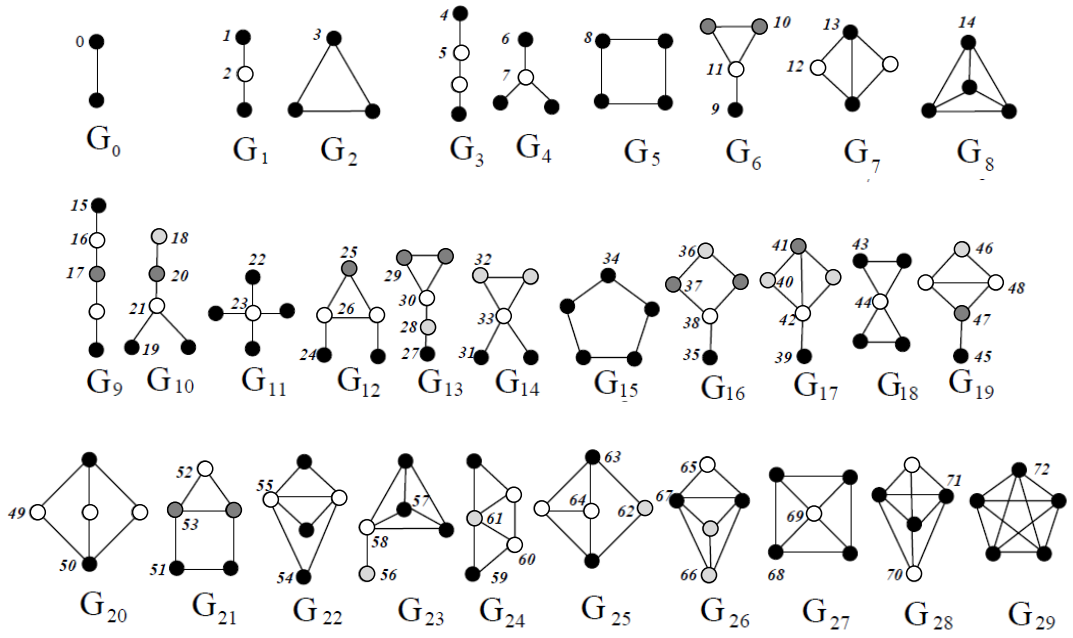


Figura 3.3 Rappresentazione dei graphlet di ordine 2 (G_0), 3 (G_1, G_2), 4 ($G_3, \ldots, G_8$) e 5 ($G_9, \ldots, G_{29}$) e delle loro orbite di automorfismo $0, 1, \ldots, 72$. In uno stesso graphlet i nodi appartenenti alla stessa orbita sono rappresentati con la stessa tonalità di grigio, distinguendo così nodi con ruolo topologico differente.

dei processi biologici che rappresentano.

I *graphlet* sono sottografi connessi e indotti della rete originaria: connessi perché per ogni coppia di nodi appartenenti ad uno stesso graphlet esiste un percorso che li unisce, indotti perché i loro nodi sono connessi mediante tutti e gli stessi archi che connettono i nodi del relativo sottografo individuato nella rete originaria. I graphlet che sono stati considerati, riportati in figura 3.3, sono quelli di ordine due, tre, quattro e cinque. La ricerca di graphlet di ordine superiore all'interno di una rete di grandi dimensioni diventa infatti un problema computazionalmente oneroso.

Dal punto di vista topologico e strutturale è particolarmente interessante distinguere i nodi che occupano una diversa posizione all'interno di uno stesso graphlet. Si osservi, ad esempio, il graphlet G_1 in figura 3.3: il nodo rappresentato in bianco ha un ruolo centrale nella topologia del graphlet, diverso dal ruolo topologico dei nodi rappresentati in nero, collocati alle estremità della struttura. La stessa cosa si può osservare per il graphlet G_4: il nodo in bianco ha un ruolo centrale nella topologia del graphlet, mentre i

tre nodi in nero occupano una posizione periferica. Per distinguere il "ruolo topologico" che ciascun nodo riveste all'interno di un graphlet in termini di relazioni con gli altri nodi del graphlet e per darne una definizione rigorosa si ricorre al concetto di *orbita di un automorfismo*.

Definizione 3.4.4 *Siano X e Y due grafi dello stesso ordine e siano E_x ed E_y rispettivamente gli archi di X e di Y. Si dice* isomorfismo tra X e Y *una funzione biunivoca $g : V(X) \mapsto V(Y)$ che mappa i vertici di X nei vertici di Y, tale che:*

$$uv \in E_x \Leftrightarrow g(u)g(v) \in E_y,$$

dove u e v sono due nodi del grafo X e uv è l'arco che li connette. L'automorfismo g è cioè tale che uv è un arco del grafo X se e solo se l'arco $g(u)g(v)$, che congiunge l'immagine dei vertici u e v secondo g, è un arco di Y.

Un *automorfismo* è un isomorfismo di un grafo su se stesso. Gli automorfismi di un grafo X formano un *gruppo*, detto *gruppo degli automorfismi di X* e indicato con $Aut(X)$.

Definizione 3.4.5 *Sia u un nodo del grafo X, si dice* orbita *dell'automorfismo di u l'insieme di nodi*

$$Orb(u) = \{v \in V(X) | v = g(u) \quad \textit{per qualche } g \in Aut(X)\},$$

dove $V(X)$ è l'insieme dei nodi di X.

L'orbita di un nodo di un grafo è cioè costituita dall'insieme dei nodi in cui il nodo stesso viene mappato da uno degli automorfismi del grafo. Secondo questa definizione è possibile tradurre in modo matematicamente rigoroso il concetto di "ruolo topologico" che i nodi rivestono all'interno di un graphlet in termini di relazioni con gli altri nodi dello stesso graphlet. Infatti i nodi con uno stesso ruolo appartengono alla stessa orbita di automorfismo o, più brevemente, alla stessa orbita. I diversi graphlet considerati, $G_0, G_1, \ldots, G_{29}$, consentono di distinguere 73 orbite, rappresentate in figura 3.3.

Per tenere in considerazione la distinzione tra le 73 diverse orbite così individuate, la *signature* di un nodo, indicata con SIG, è costituita da un vettore di 73 elementi: la

k-esima coordinata di tale vettore indica quante volte il nodo occupa, all'interno dei graphlet individuati nella rete originaria, la posizione relativa all'orbita k. Il primo elemento della signature di un nodo coincide con il grado del nodo nella rete, ovvero il numero di connessioni che il nodo presenta con altri nodi. Infatti l'unico graphlet di ordine due è il graphlet G_0, costituito da un singolo arco, e la prima coordinata della signature definisce quante volte il nodo viene "toccato" da singoli archi della rete. La signature consente dunque di caratterizzare le proprietà topologiche locali in un intorno di un nodo in termini di numero di partecipazioni ai diversi graphlet individuati nella rete, consentendo di distinguere le diverse posizioni che il nodo occupa all'interno delle strutture topologiche considerate.

La misura che si è deciso di adottare si basa sulla quantificazione delle variazioni di interazioni di ciascuna variabile genica con le restanti nelle due classi di soggetti ricorrendo ad un confronto delle signature. In particolare lo score associato a ciascun gene $g_i, i = 1, 2, \ldots, G$ si ottiene come differenza delle signature SIG_i^A e SIG_i^B dei nodi associati al gene g_i nelle due reti.

Definizione 3.4.6 *Sia g_i l'i-esima delle G variabili geniche considerate e siano SIG_i^A e SIG_i^B le signature dei nodi corrispondenti a g_i nelle due reti di interazioni geniche, A e B. Sia $D(SIG_i^A(k), SIG_i^B(k))$ la distanza tra le due signature relativa alla k-esima orbita così definita:*

$$D(SIG_i^A(k), SIG_i^B(k)) = \frac{|log(SIG_i^A(k)+1) - log(SIG_i^B(k)+1)|}{log(max\{SIG_i^A(k), SIG_i^B(k)\}+2)}. \tag{3.3}$$

Lo score basato su graphlet *associato al gene g_i è definito da:*

$$S_{i,graphlet} = \frac{\sum_{k=0}^{72} D(SIG_i^A(k), SIG_i^B(k))}{73}. \tag{3.4}$$

Le differenze tra signature dei nodi sono in scala logaritmica. Per evitare che la funzione logaritmo assuma valore infinito in corrispondenza di eventuali coordinate nulle delle signature, gli elementi delle signature dei nodi sono stati tutti incrementati di 1. L'argomento del logaritmo a denominatore nell'equazione 3.3 è stato incrementato di 2 per evitare sia che il logaritmo assuma valore infinito, sia che assuma valore nullo. Sia la distanza tra le orbite k-esime di due signature, sia lo score associato a ciascuna

Algoritmo 1 Calcolo delle signature

for $i = 1, \ldots, G$ **do**
    sia x il nodo associato al gene g_i nella rete
    $N(x) \leftarrow$ insieme dei nodi vicini di x
    for all $y \in N(x)$ **do**
        incremento di uno la conta del grado di x e y
        $N(y) \leftarrow$ insieme dei nodi vicini di y, escluso x.
        for all $z \in N(y)$ **do**
            incremento di uno la conta dell'orbita opportuna dei graphlet di ordine 3 dei
            nodi x, y e z
            $N(z) \leftarrow$ insieme dei nodi vicini di z, esclusi x e y.
            for all $k \in N(z)$ **do**
                incremento di uno la conta dell'orbita opportuna dei graphlet di ordine 4 dei
                nodi x, y, z e k
                $N(k) \leftarrow$ insieme dei nodi vicini di k, esclusi x, y e z.
                for all $w \in N(k)$ **do**
                    incremento di uno la conta dell'orbita opportuna dei graphlet di ordine 5
                    dei nodi x, y, z, k e w
                end for
            end for
        end for
    end for
end for;

variabile genica sono stati normalizzati in modo tale da ottenere valori compresi nel range $[0, 1)$. Uno score pari a 0 si ottiene quando le signature dei due nodi coincidono tra loro e la topologia locale nell'intorno dei due nodi è rimasta praticamente invariata nelle due reti. Uno score con valore prossimo a 1, invece, viene associato a geni che hanno modificato in modo significativo le loro relazioni di interazione con le altre variabili.

Gli score basati sul grado o sul coefficiente di clustering ricorrono ad una definizione della variazione di interazioni più semplice rispetto all'articolata caratterizzazione della

topologia locale delle reti di cui si avvale lo score basato su graphlet (basti pensare che il grado di un nodo è solo il primo elemento della sua signature). Nel caso dei graphlet, infatti, il confronto delle signature dei nodi delle due reti non si limita a valutare le variazioni di interazione tra coppie di geni considerate separatamente e singolarmente o tra i regolatori di uno stesso gene. La signature di un nodo dà una descrizione più complessa della partecipazione del nodo a relazioni di interazione con altri nodi della rete e il confronto delle signature consente di dare una caratterizzazione particolareggiata delle modalità di interazione del gene con le restanti variabili.

Per l'estrazione dei graphlet dalle reti costruite il software *GraphCrunch* [61] mette a disposizione una serie di strumenti per l'analisi delle proprietà topologiche globali e locali delle reti. A partire da questo software l'algoritmo per l'estrazione dei graphlet è stato riadattato per calcolare le signature dei nodi di una rete. Il metodo implementato considera uno alla volta tutti i nodi della rete e, per ciascun nodo, ripete iterativamente una serie di operazioni descritte in modo semplificato nell'Algoritmo 1. Per ciascun nodo della rete vengono presi in considerazione i nodi vicini, e i nodi vicini dei vicini e così via fino a raggiungere i nodi a distanza 4. Ad ogni iterazione per ciascun nodo preso in considerazione viene aggiornato il conteggio delle opportune orbite dei graphlet di vario ordine nelle quali il nodo è coinvolto nella rete.

3.4.4 Differenze di correlazione

L'ultima misura introdotta per la quantificazione della variazione di interazioni geniche si basa su un criterio differente rispetto alle precedenti poiché non ricorre ad una descrizione topologica delle variazioni di interazioni geniche. Si è infatti pensato di utilizzare una misura che tenga in considerazione il peso delle relazioni in base alla correlazione tra un gene e gli altri geni. Tale misura può essere direttamente applicata a partire dalle matrici dei valori di correlazione e correlazione parziale tra tutte le coppie di geni.

Per calcolare lo score da associare al gene g_i si prendono in considerazione i valori di correlazione (o correlazione parziale) tra il profilo di espressione di g_i con i profili di espressione di tutte le altre variabili geniche g_j nella classe A, r_{ij}^A, e nella classe B, r_{ij}^B,

considerati con il loro segno, e se ne calcola la differenza in modulo. Il massimo valore di differenza calcolato indica la massima variazione di correlazione tra il gene g_i e le restanti variabili nelle due classi di soggetti e viene assegnato come punteggio, dopo essere stato opportunamente normalizzato, alla variabile g_i.

Definizione 3.4.7 *Sia g_i l'i-esima delle G variabili geniche considerate e siano r_{ij}^A e r_{ij}^B con $j = 1, \dots G, j \neq i$ rispettivamente i valori di correlazione (o correlazione parziale) del profilo di espressione di g_i con i profili di espressione dei restanti geni g_j nelle due classi di soggetti A e B. Lo* score basato sulla variazione di correlazione *associato al gene g_i è definito da:*

$$S_{i,corr.} = \frac{\max_{j \neq i} |r_{ij}^A - r_{ij}^B|}{2}. \tag{3.5}$$

In questo caso è stata effettuata una normalizzazione rispetto al massimo valore di variazione della correlazione (o correlazione parziale) che è pari a 2 poiché sia la correlazione che la correlazione parziale assumono valori nell'intervallo $[-1, 1]$.

Rispetto alle precedenti questa misura consente di tenere in considerazione le variazioni di interazione tra geni che si traducono in un cambiamento di segno della correlazione o della correlazione parziale. Se la correlazione tra due variabili geniche subisce una variazione di segno nelle due classi A e B, non è detto che questo comporti una variazione della topologia delle due reti. Infatti, se i valori di correlazione sono entrambi elevati (piccoli) in modulo, pur avendo segno opposto, l'arco che congiunge le due variabili verrà introdotto (non sarà presente) in entrambe le reti costruite. Di conseguenza questa variazione della relazione tra le due variabili geniche non può essere rilevata osservando esclusivamente la topologia delle reti costruite. É invece più facile rilevarla considerando la differenza dei valori di correlazione che assumerà con buona probabilità un valore elevato quando i valori di correlazione hanno segno opposto.

Tuttavia una limitazione di questa misura rispetto a quelle basate sul confronto della topologia delle reti costruite è legata al fatto che essa tiene in considerazione l'interazione dei geni con uno soltanto degli altri geni, ovvero quella che ha subito una massima variazione, e non consente una caratterizzazione "globale" della variazione delle modalità di interazione con tutte le variabili.

4

Dati simulati

Prima dell'applicazione del nuovo metodo di analisi proposto per la selezione robusta di biomarcatori a dati reali, questo è stato applicato a dati di espressione genica simulati. Nella prima parte del capitolo viene presentato il simulatore utilizzato descrivendone le caratteristiche che lo rendono un valido strumento per la simulazione di dati di espressione genica. Nella seconda parte, invece, sono riassunte le fasi della simulazione che hanno consentito di ottenere i dati di espressione genica di due diverse classi di popolazione di soggetti.

4.1 Obiettivi della simulazione

I metodi di analisi proposti, descritti nel capitolo 3, sono pensati per essere applicati a dati di espressione genica ottenuti mediante esperimenti con microarray. L'applicazione di tali metodi consente, in particolare, di mettere a confronto dati di espressione statici relativi a due classi di popolazione di soggetti. Possono essere messi a confronto non solo soggetti sani e soggetti malati, ma anche soggetti che presentano due diverse patologie o forme diverse della stessa patologia o ancora soggetti che presentano la stessa patologia ma a diversi stadi della sua evoluzione.

L'estrazione di informazioni dai dati e il metodo di elaborazione proposto hanno come fine ultimo quello di proporre un criterio per individuare i geni cosiddetti biomarcatori. Come già precisato nel capitolo precedente, sono definiti biomarcatori quei geni la cui attività e le cui relazioni di interazione con altri geni vengono alterate durante l'evoluzione di una patologia. La possibilità di riconoscere e individuare tali geni è di fondamentale importanza sia per la diagnosi che per la terapia di patologie complesse.

Prima di applicare il nuovo metodo proposto a dati reali di espressione genica, è opportuno procedere ad una fase di validazione mediante l'utilizzo di dati di espressione simulati. Ricorrendo alla simulazione dei profili di espressione, infatti, si producono dei dati rappresentativi di una situazione completamente nota: è possibile quindi stabilire

quali siano i geni la cui attività viene alterata tenendo in considerazione le relazioni che intercorrono tra i geni simulando una rete di regolazione nota [23]. In questo modo la lista dei geni biomarcatori è nota ed è possibile valutare il livello di precisione con cui il metodo proposto consente di individuare tali biomarcatori.

L'obiettivo della simulazione è, naturalmente, quello di produrre dei dati con caratteristiche simili a quelle dei profili di espressione genica reali. Nel lavoro presentato, sono stati simulati dati di espressione relativi a due popolazioni di soggetti che possono essere pensati come appartenenti a due classi diverse di una stessa patologia. Nel processo di simulazione, descritto nel dettaglio nel seguito del capitolo, è stata riprodotta la variabilità biologica che caratterizza i soggetti reali, simulando anche l'eterogeneità presente tra i soggetti appartenenti alla stessa classe. Le due classi della patologia, inoltre, sono state riprodotte definendo diversi insiemi di geni la cui attività viene compromessa dalla malattia.

4.2 Il simulatore

Per generare i profili di espressione genica è stato utilizzato un simulatore di reti di interazione genica, descritto nel dettaglio in [23], capace di riprodurre le caratteristiche principali delle reti di regolazione trascrizionale reali. Per simulare i valori di espressione dei geni che prendono parte ai complessi meccanismi di regolazione e controllo cui si è fatto cenno nella sezione 2.1 è infatti necessario, in primo luogo, simulare la rete di regolazione che definisce le interazioni tra i geni e la loro azione regolatoria.

Il simulatore riproduce, come schematizzato in figura 4.1, le proprietà delle reti di regolazione reali legate alla topologia della rete, ai meccanismi di interazione tra geni e alla dinamica dell'espressione genica. In particolare la topologia della rete viene generata cercando di riprodurre alcune delle proprietà topologiche delle reti reali, quali ad esempio una distribuzione di tipo *scale-free* per il grado di connettività dei nodi e un *coefficiente di clustering* costante indipendentemente dal numero di nodi della rete. Le interazioni tra i regolatori della trascrizione di ciascun gene vengono rappresentate mediante *logica fuzzy*. Infine il livello di espressione di ciascun gene viene

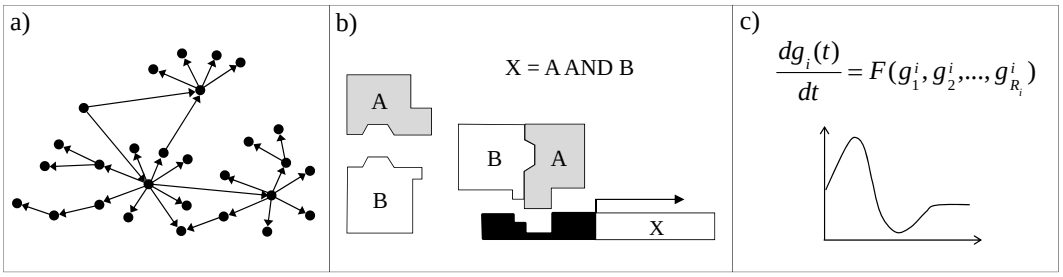


Figura 4.1 (a) Il simulatore riproduce la topologia delle reti di regolazione, (b) rappresenta le interazioni tra regolatori mediante funzioni logiche, (c) descrive la dinamica di espressione mediante equazioni differenziali.

descritto mediante l'utilizzo di equazioni differenziali che consentono di rappresentare la complessa dinamica dei meccanismi di espressione.

4.2.1 Topologia della rete di regolazione

La topologia della rete di regolazione viene generata organizzando i geni secondo una struttura gerarchica su più livelli che consente di riprodurre tre delle principali proprietà topologiche esibite dalle reti di regolazione reali.

In primo luogo viene riprodotta l'organizzazione di tipo scale-free tipicamente esibita dalle reti trascrizionali reali [62]. Il numero di connessioni presenti nella rete segue una distribuzione di tipo *power-law*: la probabilità che un nodo abbia grado di connettività K è data da

$$P(K) \propto \frac{1}{K^{\gamma}}, \tag{4.1}$$

dove γ è un parametro caratteristico della distribuzione, con valore prossimo a 2.2 nelle reti reali [63]. Più precisamente la distribuzione del numero di archi entranti nei nodi della rete, detta *in-degree distribution*, è di tipo scale-free, mentre il numero di archi uscenti da ciascun nodo segue una distribuzione, *out-degree distribution*, che può essere di tipo scale-free o esponenziale, a seconda dell'organismo [64].

Un'altra proprietà topologica che caratterizza le reti biologiche è rappresentata dal coefficiente di clustering, che descrive la tendenza dei nodi connessi ad uno stesso nodo g_i a formare connessioni tra di loro. Tale parametro, secondo la definizione 3.4.2, è

dato dal rapporto tra il numero di connessioni esistenti tra i vicini di uno stesso nodo e il numero totale di possibili connessioni che questi potrebbero formare tra loro. E' stato verificato che, nelle reti biologiche, il coefficiente di clustering medio C di tutti i nodi assume un valore costante indipendentemente dal numero di nodi presenti nella rete [65].

L'ultima proprietà topologica delle reti biologiche presa in considerazione è la distanza media L tra tutte le coppie di nodi della rete. Si è osservato che le reti trascrizionali reali sono caratterizzate da un valore relativamente basso di L, tipicamente inferiore a 5, evidenziando un'organizzazione della rete di tipo *small-world* [66].

I modelli più frequentemente utilizzati per caratterizzare la topologia delle reti di regolazione non sono in grado di riprodurre contemporaneamente tutte e tre le proprietà topologiche riportate. I modelli random [67], in cui ciascun nodo ha eguale probabilità di essere connesso a un qualsiasi altro nodo della rete, sono in grado di riprodurre esclusivamente l'organizzazione di tipo small-world. I modelli di tipo scale-free [68] riproducono correttamente la distribuzione del grado di connettività dei nodi e l'organizzazione small-world, tuttavia definiscono delle reti caratterizzate da un coefficiente di clustering che decresce con il numero di nodi. Infine i modelli di tipo geometrico [69], introdotti poiché consentono di costruire reti con coefficiente di clustering costante indipendentemente dal numero di nodi, non esibiscono una distribuzione di tipo power-law rispetto al grado di connettività dei nodi.

Nessuno dei tre modelli mostra simultaneamente le tre proprietà topologiche riscontrate nelle reti reali. Un modello di tipo gerarchico delle reti è stato proposto in [65], in grado di riprodurre le caratteristiche della distribuzione scale-free e di un coefficiente di clustering elevato indipendentemente dal numero di nodi che compongono la rete. A partire da tale modello, il simulatore adottato riproduce la topologia delle reti di regolazione ricorrendo ad un modello gerarchico modulare (*modular topology model*): diverse tipologie di moduli di interazione tra geni vengono replicate, con un certo grado di flessibilità, sui diversi livelli in cui viene organizzata la rete. In particolare la rete viene costruita secondo una procedura iterativa che consiste in tre passi principali:

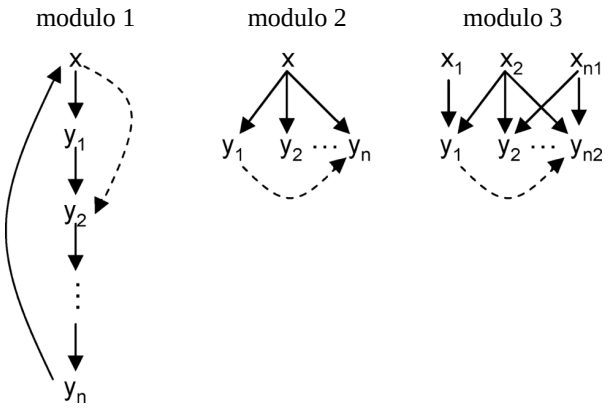


Figura 4.2 Strutture modulari utilizzate per riprodurre la topologia della rete. Con la lettera x sono indicati i geni che hanno il ruolo di regolatori (o regolatori principali).

1. Vengono generati tre moduli con un numero random di nodi e struttura parzialmente casuale.

2. Ad ognuno dei tre moduli viene assegnato un punteggio e viene campionato un modulo con probabilità proporzionale a tale punteggio. Sia m il numero di nodi del modulo selezionato.

3. Vengono scelti, tra i nodi della rete, m nodi che vengono connessi tra loro secondo la struttura topologica del modulo selezionato.

Alla prima iterazione i tre passi vengono ripetuti finché tutti i nodi della rete appartengono ad uno dei moduli introdotti e la rete è costituita da un numero di strutture modulari tra loro non connesse. Nelle iterazioni successive i moduli vengono utilizzati per connettere tra loro, ripetendo i tre passi, non più i singoli nodi della rete ma le strutture modulari costruite all'iterazione precedente. In questo modo la topologia della rete ottenuta è caratterizzata da strutture modulari diverse e di diversa dimensione che si ripropongono su più livelli.

Le strutture modulari utilizzate, riprodotte in figura 4.2, rappresentano le diverse tipologie di moduli caratterizzanti le interazioni tra geni nella rete di regolazione. Tali strutture sono state individuate nelle reti trascrizionali reali, dove si presentano con una frequenza significativamente superiore rispetto a quella con cui si presentano nelle reti random [70]. La possibilità di far variare il numero di nodi di ciascun modulo

conferisce un certo grado di flessibilità alla topologia riprodotta. Inoltre ciascun modulo presenta la possibilità di introdurre in maniera random alcune connessioni tra i nodi che lo compongono (rappresentate da una linea tratteggiata in figura 4.2), oltre a quelle prefissate (rappresentate da una linea continua), in modo da ottenere il coefficiente di clustering desiderato. Infine, sia la scelta del modulo da introdurre nella rete, sulla base del punteggio assegnato al passo 2, sia la scelta dei nodi che verranno connessi tra loro nella rete sono operate in modo tale da garantire una distribuzione di tipo power-law del grado di connettività dei nodi.

4.2.2 Interazioni tra regolatori

Una volta simulata la topologia della rete di regolazione, ciascun nodo della rete g_i è caratterizzato da R_i archi entranti che lo connettono con i suoi regolatori $g_1^i, g_2^i, \ldots, g_{R_i}^i$. Per poter simulare il profilo di espressione del gene g_i, determinandone il livello di attivazione in ciascun istante temporale t, è necessario in primo luogo caratterizzare i meccanismi con cui interagiscono i suoi regolatori per attivarne o inibirne la trascrizione. Il livello di espressione $T_i(t)$ cui il gene g_i tende all'istante temporale t viene dunque espresso in funzione del livello di espressione dei suoi regolatori:

$$T_i(t) = F(g_1^i, g_2^i, \ldots, g_{R_i}^i).$$

Il livello di espressione di ciascun regolatore viene normalizzato rispetto al suo livello di espressione massimo. Di conseguenza anche $T_i(t)$ rappresenta il livello di espressione normalizzato cui il gene g_i tende, assumendo valori nell'intervallo $[0, 1]$.

L'azione combinata dei regolatori che interagiscono tra loro per controllare e determinare il livello di espressione di g_i può essere efficacemente rappresentata mediante l'utilizzo delle funzioni logiche AND, OR, NOT. Ad esempio, la funzione AND può essere utilizzata per rappresentare un'azione regolatoria di tipo cooperativo, il cui effetto si ottiene soltanto quando tutti i regolatori sono contemporaneamente attivi. La funzione NOT invece può essere utilizzata per rappresentare un'azione di tipo inibitorio sulla trascrizione di g_i da parte dei suoi regolatori.

Le funzioni logiche, tuttavia, possono essere applicate esclusivamente a variabili di tipo binario e non consentono quindi di tenere in considerazione la natura intrinseca-

mente continua dei meccanismi di regolazione genica. Per risolvere questa limitazione è possibile descrivere le interazioni tra regolatori ricorrendo a funzioni basate sulla logica fuzzy, che rappresenta un'estensione della logica Booleana ad un dominio continuo. In particolare il simulatore adottato definisce quattro diverse funzioni, basate su logica fuzzy, per caratterizzare l'interazione tra regolatori:

- $f_{\text{COOPERATIVA}}(g_1^i, g_2^i, \ldots, g_{R_i}^i) = min\{g_1^i, g_2^i, \ldots, g_{R_i}^i\}$, descrive un'azione regolatoria attuata solo se tutti i regolatori sono contemporaneamente attivi;

- $f_{\text{SINERGICA}}(g_1^i, g_2^i, \ldots, g_{R_i}^i) = min\{1, sum\{g_1^i, g_2^i, \ldots, g_{R_i}^i\}\}$, descrive un'azione regolatoria attuata da diversi regolatori, anche se non contemporaneamente attivi;

- $f_{\text{NEGATIVA}}(g_r^i) = 1 - g_r^i$, descrive un effetto inibitorio della trascrizione da parte del regolatore r;

- $f_{\text{COMPETITIVA}}(g_r^i, g_{r'}^i) = max\{0, g_{r'}^i - g_r^i\}$, descrive un effetto inibitorio della trascrizione attuato da due regolatori, r e r', tra loro in competizione.

Inoltre, per modulare la diversa efficienza di regolazione, i valori di espressione dei regolatori di ciascun gene g_i vengono pesati utilizzando i parametri $w_{ir} \geq 0$. Valori elevati di w_{ir} sono indicativi di un'intensa azione regolatoria del gene g_r^i sul gene g_i.

In conclusione, il livello di espressione normalizzato $T_i(t)$ calcolato per ciascun gene g_i è il risultato della combinazione delle funzioni logiche definite, applicate ai regolatori opportunamente scalati mediante i w_{ir}. In questo modo è possibile descrivere varie interazioni regolatorie caratterizzate da diverso grado di complessità.

4.2.3 Dinamica dell'espressione genica

Nelle reti biologiche reali i meccanismi di regolazione dell'espressione genica non sono istantanei, bensì caratterizzati da una complessa dinamica. A tal proposito vengono utilizzate le equazioni differenziali che permettono di descrivere l'evoluzione dei processi di trascrizione e di degradazione di ciascun prodotto genico. La possibilità di introdurre parametri specifici per ciascun gene consente di descrivere numerosi fenomeni che

caratterizzano la cinetica dell'espressione genica, come ad esempio la presenza di ritardi nella trascrizione, fenomeni di saturazione che intervengono durante la regolazione o, ancora, l'effetto di una soglia di attivazione qualora la trascrizione fosse attivata solo in presenza di una certa concentrazione di complessi regolatori. Infine, variando opportunamente le condizioni iniziali, è possibile caratterizzare diversi stimoli esterni che perturbano il sistema.

Per tenere conto di fenomeni di saturazione e della presenza di una soglia di attivazione della trascrizione, il livello di espressione $T_i(t)$ può essere modulato mediante la seguente funzione di attivazione sigmoidale:

$$S_i(T_i(t), \alpha_i, \beta_i) = \frac{1}{1 + exp[-\alpha_i(T_i(t) - \beta_i)]}, \tag{4.2}$$

dove α_i e β_i sono parametri specifici del gene g_i. Infine la velocità di variazione del livello di espressione genica del gene g_i al tempo t viene descritta mediante l'equazione differenziale:

$$\frac{dg_i(t)}{dt} = \lambda_i[S_i(T_i(t), \alpha_i, \beta_i) - g_i(t)], \tag{4.3}$$

dove λ_i è una costante temporale che modula l'effetto combinato dei processi trascrizionali e di degradazione di g_i.

L'integrazione di logica fuzzy ed equazioni differenziali nel processo di simulazione consente di ottenere una rete di regolazione caratterizzata da un elevato livello di complessità. Le funzioni logiche caratterizzano nel dettaglio le interazioni tra i regolatori della trascrizione, di cui le equazioni differenziali non tengono conto. Al contrario le equazioni differenziali sono in grado di descrivere la complessa dinamica della regolazione che non può essere caratterizzata ricorrendo unicamente all'utilizzo delle funzioni logiche. Il simulatore così descritto consente dunque di ottenere una varietà di profili di espressione genica caratterizzati da una dinamica con proprietà e complessità confrontabili con quelle dei profili di espressione reali.

4.3 Simulazione dei dati

Utilizzando il simulatore descritto nella sezione precedente è stato possibile ottenere i profili di espressione di due classi di soggetti, ciascuna costituita da $M = 500$ campioni,

supponendo di monitorare $G = 10000$ geni. Ciascun soggetto è caratterizzato da una matrice di connettività W, che definisce la topologia della rete di regolazione, e da un vettore $x_n = (x_{1n}, \ldots, x_{Gn})$ contenente i valori di espressione dei geni monitorati per il soggetto $n = 1, \ldots, N$.

La matrice di connettività W contiene i parametri utilizzati per modulare la diversa efficienza di regolazione dei regolatori di uno stesso gene. Gli elementi w_{ij} sono diversi da zero in presenza di un'azione regolatoria del gene g_j sul gene g_i: il segno e il valore di w_{ij} indicano il segno e l'intensità dell'azione regolatoria. La matrice di connettività W può essere pensata come rappresentativa del genotipo di ciascun soggetto. Infatti, i principali meccanismi di controllo e regolazione dell'espressione genica sono quelli che si esplicano a livello del controllo trascrizionale. Ciascuna sequenza genica è preceduta da una sequenza nucleotidica, detta promotore, alla quale si legano le proteine che ne controllano la trascrizione (fattori di trascrizione), attivandola, inibendola o regolando velocità ed efficienza del processo. Il parametro w_{ij} può dunque essere interpretato come l'affinità tra il promotore del gene g_i e le proteine e i fattori di trascrizione che, codificati dal gene g_j, ne regolano la trascrizione. Un diverso valore di w_{ij} nelle matrici di connettività di due soggetti indica una diversa affinità tra il promotore di g_i e i fattori trascrizionali. Tale differenza di affinità è tipicamente legata ad una differenza nelle sequenze nucleotidiche del promotore ed è dunque riconducibile ad una differenza genotipica tra i due soggetti.

Il vettore x_n, invece, contiene i valori di espressione dei G geni allo stato stazionario, ottenuti risolvendo le equazioni differenziali della dinamica (4.3). In particolare l'i-esimo elemento del vettore x_n rappresenta il livello di espressione del gene g_i allo stato stazionario, calcolato risolvendo l'equazione differenziale (4.3) relativa al gene g_i a partire da prefissate condizioni iniziali. Il vettore di espressione x_n può essere pensato come rappresentativo del fenotipo di ciascun soggetto in una particolare condizione ambientale (determinata dalle condizioni iniziali imposte alle equazioni differenziali).

Per quanto riguarda la definizione della topologia delle reti di regolazione simulate è stato adottato il procedimento descritto nella sezione 4.2.1, basato su un modello gerarchico e modulare della topologia. La topologia ottenuta è dunque caratterizzata da

una distribuzione di tipo scale-free del grado di connettività dei nodi, da un coefficiente di clustering indipendente dal numero di nodi e da un'organizzazione strutturale di tipo small-world. La topologia della rete di regolazione è la stessa per tutti i soggetti simulati. Ciò che varia, al fine di riprodurre la variabilità genotipica dei soggetti, è la matrice W che definisce i pesi w_{ij} associati a ciascun arco della rete di regolazione.

I profili di espressione x_n associati a ciascun soggetto sono invece stati calcolati risolvendo le equazioni differenziali (4.3) per ciascun gene e per ciascun soggetto, partendo dalle stesse condizioni iniziali. I parametri α_i e β_i che caratterizzano la dinamica di espressione di ciascun gene g_i sono stati campionati da distribuzioni Gaussiane con media μ_α, μ_β e deviazione standard σ_α, σ_β. I valori medi μ_α e μ_β sono stati impostati rispettivamente a 20 e 0.2. I valori di σ_α e σ_β sono invece stati a loro volta campionati da due diverse distribuzioni Gaussiane: $\sigma_\alpha \sim \mathcal{N}(0.5, 0.075)$ e $\sigma_\beta \sim \mathcal{N}(0.02, 0.0025)$.

4.3.1 Simulazione della variabilità di popolazione

Nella prima fase di simulazione viene generata un'unica popolazione di 1000 soggetti che presentano un certo grado di eterogeneità. Per ogni soggetto è stata riprodotta la variabilità biologica introducendo una variabilità genotipica attraverso la variazione dei pesi presenti nella matrice di connettività associata W, rappresentativa del genotipo del soggetto.

Per simulare tale varibilità si parte da una popolazione iniziale di 1000 soggetti, ottenuta a partire da uno stesso progenitore. La matrice di connettività del progenitore W^P, generata dal simulatore, è caratterizzata da quattro reti, tre delle quali costituite da 300 geni ciascuna e la quarta costituita dai restanti 9100 geni. I parametri α_i^P e β_i^P, che caratterizzano la dinamica di espressione del progenitore, sono stati campionati come precedentemente precisato. I soggetti della popolazione iniziale hanno tutti la stessa matrice di connettività del progenitore W^P, e dunque lo stesso genotipo. Tuttavia si differenziano in termini di fenotipo, poiché variano i parametri della dinamica, campionati, per ciascun soggetto, dalle distribuzioni Gaussiane: $\alpha_i \in \mathcal{N}(\alpha_i^P, 0.075)$ e $\beta_i \in \mathcal{N}(\beta_i^P, 0.0025)$.

La popolazione iniziale viene poi fatta evolvere per un certo numero di generazioni successive. Ad ogni generazione ciascun soggetto della progenie è caratterizzato da una matrice di connettività ottenuta combinando quelle dei due genitori, appartenenti alla generazione precedente, e introducendo alcune mutazioni nei pesi w_{ij}. Le mutazioni nella matrice di connettività W sono state introdotte esclusivamente nella sottomatrice relativa alle tre reti da 300 nodi ciascuna, indicata con W_{900}. La variabilità nei parametri della dinamica viene introdotta con la stessa descrizione statistica utilizzata per la popolazione iniziale.

Per ciascuno dei nuovi soggetti generati è necessario definire un limite al grado di variabilità che si è disposti ad introdurre al fine di garantirne la sopravvivenza. A tale scopo, il fenotipo con maggiore probabilità di garantire la sopravvivenza di un soggetto viene identificato con il fenotipo medio della popolazione di progenitori. Ciascun nuovo soggetto generato sopravvive solo se il suo fenotipo non si discosta troppo dal fenotipo iniziale.

In particolare, le generazioni in cui evolve la popolazione iniziale si ottengono ripetendo iterativamente i seguenti passi:

1. *Pairing:* ciascun soggetto della progenie viene creato scegliendo in modo casuale due genitori tra i soggetti della generazione precedente; la matrice di connettività W_{900} si ottiene combinando le righe delle matrici di connettività dei genitori scegliendole con eguale probabilità dalla matrice dell'uno o dell'altro genitore.

2. *Mutation:* le mutazioni genetiche vengono simulate modificando ciascun elemento w_{ij} diverso da zero con probabilità pari a $0.025/N_w$, dove $N_w = 1619$ è il numero di elementi non nulli di W_{900}. In questo modo ciascun nuovo soggetto generato ha probabilità pari a 0.025 di subire almeno una mutazione. Il nuovo valore da attribuire ai w_{ij} viene campionato da una distribuzione Gaussiana con media nulla e varianza unitaria.

3. *Selection:* tra tutti i soggetti generati che presentano la mutazione di almeno uno degli elementi w_{ij} vengono selezionati soltanto quelli in grado di sopravvivere. Sia d la distanza Euclidea tra il profilo di espressione di ciascun soggetto e il profilo di espressione medio della popolazione iniziale, ovvero il fenotipo con

maggior probabilità di sopravvivenza. Se la distanza d è inferiore ad una "soglia di sopravvivenza" pari a 0.276 (valore corrispondente al percentile 99.5 delle distanze osservate) il nuovo soggetto sopravvive, altrimenti viene eliminato.

Ad ogni generazione vengono prodotti 1000 nuovi soggetti, indipendentemente dal numero di soggetti sopravvissuti nella generazione precedente. L'evoluzione procede per 150 generazioni, fino a quando si ottiene una popolazione finale costituita da 1000 soggetti caratterizzati da genotipo diverso e da fenotipo adatto alla sopravvivenza.

4.3.2 Simulazione della malattia

Una volta riprodotta la popolazione di 1000 soggetti caratterizzati dalla variabilità genotipica descritta nella sezione precedente, per portare a termine il processo di simulazione è necessario definire e riprodurre sui soggetti simulati la condizione patologica desiderata. Essendo nota a priori la rete di regolazione genica, è possibile riprodurre le caratteristiche di quelle patologie che, causate da mutazioni a livello del genoma, compromettono l'attività di alcuni geni e alterano una parte dei meccanismi di regolazione dell'espressione, mimando alterazioni su pathway di geni, caratteristiche di patologie complesse come il cancro.

In particolare è possibile, ad esempio, silenziare completamente o ridurre il livello di espressione dei geni bersaglio della malattia. Modifiche di questo tipo alterano il profilo di espressione dei geni bersaglio, ma si ripercuotono anche sui profili di espressione di tutti quei geni che, nella rete, sono direttamente o indirettamente regolati dai geni bersaglio. Il silenziamento di un gene prende il nome di *knock-out*, mentre si indica come *knock-down* una riduzione del livello di espressione di un gene.

In fase di simulazione il knock-out del gene g_j si ottiene in primo luogo annullando il suo livello di espressione iniziale, ovvero ponendo a zero le condizioni iniziali dell'equazione differenziale che ne descrive la dinamica di espressione. In secondo luogo si annullano tutti gli elementi della j-esima riga della matrice di connettività W, annullando l'azione di inibizione/attivazione esercitata dai regolatori di g_j sulla dinamica di espressione di g_j. In formule:

$$\text{knock-out di } g_j: \qquad g_j(0) \leftarrow 0; \quad w_{ji} \leftarrow 0 \quad \forall i = 1, \ldots, G.$$

Gene	Out-degree Normalizzato	Stato Stazionario
221	1	0.9994
869	0.87	0.9897
114	0.73	0.9994
429	0.73	0.8843
612	0.67	0.9993
417	0.6	0.6
20	0.53	0.7666
3	0.53	0.7667
323	0.53	0.7667
752	0.53	0.7666
595	0.47	0.7333
761	0.47	0.7333

Tabella 4.1 Hub delle due malattie: nelle prime sei righe sono riportati gli hub della malattia A, nelle ultime sei righe, invece, sono indicati i 6 hub aagiuntivi considerati per simulare la malattia B.

Il knock-down del gene g_j si ottiene, invece, dimezzando sia il suo livello di espressione iniziale sia l'azione regolatoria dei suoi regolatori. In formule:

$$\text{knock-down di } g_j: \qquad g_j(0) \leftarrow g_j(0)/2; \quad w_{ji} \leftarrow w_{ji}/2 \quad \forall i = 1, \ldots, G.$$

Per procedere alla simulazione delle due diverse malattie, in primo luogo i 1000 soggetti simulati sono stati suddivisi in due gruppi da 500 soggetti ciascuno. Le due malattie simulate si distinguono in base ai geni bersaglio la cui attività viene alterata. Per simulare la malattia A è stato definito un insieme di 6 geni bersaglio: il primo gruppo di 500 soggetti è stato sottoposto a knock-out/knock-down di questi 6 geni, con le modalità descritte in seguito, ottenendo la prima classe di 500 soggetti affetti dalla malattia A. I geni bersaglio della malattia B comprendono altri 6 geni oltre a quelli che caratterizzano la malattia A. Il secondo gruppo di 500 soggetti afferenti alla classe della malattia B è caratterizzato dall'ulteriore knock-out/knock-down di questo secondo insieme di geni bersaglio. I 500 soggetti di quest'ultima classe, hanno dunque subito, complessivamente, knock-out/knock-down di 12 geni.

I geni bersaglio delle due malattie sono stati scelti tra i geni con elevato *out-degree*

(ovvero con elevato numero di archi uscenti nella rete di regolazione) e con elevato livello di espressione allo stato stazionario. Per queste proprietà i geni bersaglio vengono definiti "*hub* della malattia". I geni bersaglio sono stati scelti con queste caratteristiche in modo tale che il loro knock-out (o knock-down) produca effetti sensibili sulla dinamica di espressione complessiva dei G geni simulati. Nella tabella 4.1 sono elencati i geni bersaglio scelti per le due malattie.

Per riprodurre l'eterogeneità nei meccanismi di alterazione che caratterizzano i soggetti della stessa classe di malattia, viene effettuato, per la malattia A, il knock-out/knock-down di 4, 5 o 6 geni bersaglio. La proporzione di soggetti con 4, 5 o 6 geni alterati è pari a $1/3$, $1/3$, $1/3$ rispettivamente. Ciascun hub ha probabilità $1/3$ di subire knock-out e probabilità $2/3$ di subire knock-down del suo livello di espressione.

Per quanto riguarda la malattia B, oltre ai 6 geni selezionati sono stati considerati altri 6 geni seguendo lo stesso criterio di selezione precedentemente descritto. La malattia B è dunque caratterizzata complessivamente da 12 hub. Per riprodurre l'eterogeneità della malattia il knock-out/knock-down del secondo gruppo di 6 hub è stato eseguito sui soggetti della classe B con le stesse modalità riportate per la classe A. In particolare, dunque, ciascun soggetto della classe B può subire da 4 a 6 knock-out (o knock-down) nel primo gruppo di 6 hub e da 4 a 6 knock-out (o knock-down) applicati al secondo gruppo di hub. La proporzione di knock-out e knock-down è pari rispettivamente a $1/3$ e $2/3$ come nel caso della malattia A.

4.3.3 Scalatura dei dati e aggiunta di rumore

Poiché il simulatore produce dati normalizzati, i profili simulati sono costituiti da valori di espressione compresi nell'intervallo $[0, 1]$. Per ricondurre i valori di espressione a range simili a quelli dei dataset reali è stata applicata una scalatura dei dati. I valori di espressione di ciascun gene per ciascun soggetto sono stati scalati applicando la seguente formula:

$$g_i^{new} = g_i^{old}\Delta + m - \frac{\Delta}{2};$$

dove m è la media e Δ è la differenza tra il massimo e il minimo dei valori di espressione di un dataset reale di soggetti affetti da carcinoma mammario. Il dataset è pubblico e

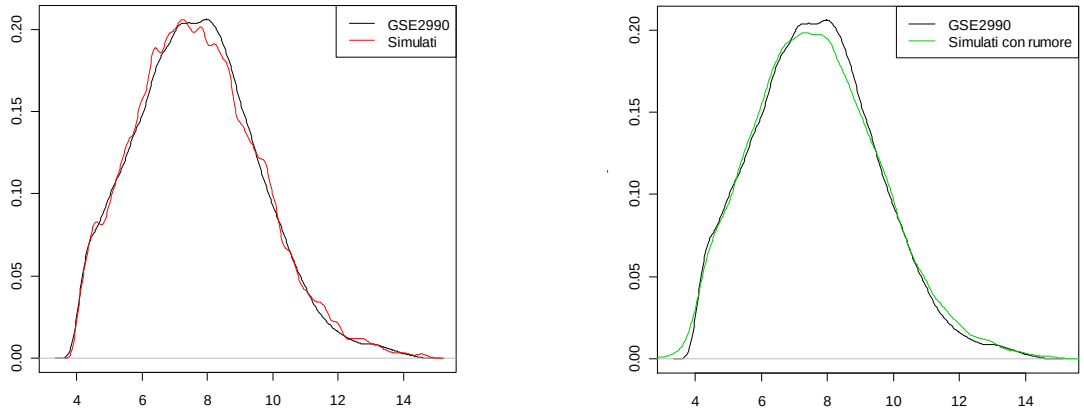


Figura 4.3 Distribuzione dei dati simulati e dei dati reali a confronto: a sinistra è riportata la distribuzione dei dati simulati scalati senza rumore, a destra invece è riportata la distribuzione dei dati simulati scalati e con aggiunta di rumore.

reperibile dal database GEO (*Gene Expression Omnibus*) [71] con codice identificativo $GSE2990$.

Ai dati scalati così ottenuti è stato aggiunto del rumore, descritto da una distribuzione Gaussiana con media nulla e varianza caratterizzata secondo una distribuzione lognormale con media 0.22 e deviazione standard 0.35. La distribuzione dei dati simulati scalati e con aggiunta di rumore corrisponde con buona approssimazione alla distribuzione dei dati di espressione reali (Wilkoxon test, p-value $= 0.9$), come si può osservare in figura 4.3.

4.3.4 Definizione dei biomarcatori

Il vantaggio principale legato all'utilizzo di dati di espressione simulati, è dato dalla possibilità di conoscere a priori la lista dei geni biomarcatori. Disponendo di tale informazione è infatti possibile quantificare la precisione del metodo di analisi proposto nel selezionare tali geni.

In fase di simulazione sono stati definiti come biomarcatori quei geni che consentono di distinguere i soggetti della classe A da quelli della classe B, ovvero quei geni il cui livello di espressione allo stato stazionario risulta significativamente diverso per i soggetti nelle due classi A e B. Per definire la lista dei biomarcatori sono stati considerati

Livello	Numero di biomarcatori
l_0	6
l_1	35
l_2	41
l_3	45
l_4	23
l_5	15
l_6	9
l_7	7
l_8	4

Tabella 4.2 Biomarcatori suddivisi per livello in base alla loro distanza dai geni G^B_{KOKD}.

i profili di espressione statici dei 500 soggetti della classe A e dei 500 soggetti della classe B. É stata creata una matrice che contiene, per ciascuno dei 10000 geni considerati, le differenze del livello di espressione tra soggetti della classe A e soggetti della classe B. Da questa matrice è stata poi calcolata, per ciascun gene, la mediana delle differenze dei valori di espressione nelle due classi. Sono stati considerati biomarcatori i geni la cui mediana delle differenze è superiore a 10^{-6}.

In questo modo è stata ottenuta una lista di 185 biomarcatori, tra i quali ricadono i 6 geni che hanno subito knock-out/knock-down nei soggetti della classe B ma non in quelli della classe A, indicati con G^B_{KOKD}. Questi geni, infatti, subiscono un'alterazione del livello di espressione nei soli soggetti che appartengono alla classe B. Tra i biomarcatori sono inoltre compresi tutti quei geni il cui livello di espressione subisce in qualche modo gli effetti del knock-out/knock-down dei geni G^B_{KOKD}: si tratta dei geni che, nella rete di regolazione simulata, sono direttamente o indirettamente regolati dai geni G^B_{KOKD}.

I biomarcatori così definiti possono essere suddivisi per livelli in base alla loro distanza dai geni G^B_{KOKD} nella rete di regolazione. In particolare i biomarcatori appartenenti al livello l_i sono quelli che si trovano a distanza i dai geni G^B_{KOKD}. In tabella 4.2 è riportato il numero di geni biomarcatori per ciascun livello. I biomarcatori appartenenti al livello l_0 coincidono con i geni G^B_{KOKD} mentre i biomarcatori più lontani dai geni G^B_{KOKD} nella rete di regolazione si trovano a distanza 8.

5

Analisi dei risultati

I nuovi metodi per il ranking dei geni e l'identificazione dei biomarcatori sono stati applicati ai dati di espressione simulati, descritti nel capitolo 4, relativi a due classi di soggetti. Dopo una prima verifica sulla confrontabilità delle reti di interazioni geniche costruite per le due classi di soggetti in base alle loro proprietà topologiche globali, è stata valutata la capacità di individuare i geni biomarcatori per i metodi proposti, confrontandoli con il metodo classico di selezione basato su test SAM. Nell'ultima parte del capitolo viene proposta un'analisi delle prestazioni dei metodi di ranking e della riproducibilità dei risultati ottenuti in corrispondenza di una riduzione del numero di campioni disponibili per le variabili geniche monitorate.

5.1 Proprietà delle reti di interazioni geniche identificate

La prima fase di elaborazione dei dati consiste nella costruzione delle reti di interazioni geniche per due classi di soggetti A e B, analizzando i dati di espressione simulati. Le reti sono state costruite mediante confronto pair-wise dei profili di espressione genica, come descritto nel capitolo 2, utilizzando come misure di similarità la correlazione e la correlazione parziale.

Per una prima validazione dei metodi proposti, l'analisi è stata ristretta alle prime 900 variabili geniche simulate. Si tratta dei 900 geni corrispondenti alle 3 componenti della rete di regolazione simulata costituite da 300 geni ciascuna e rappresentate dalla matrice di connettività W_{900}. Questi 900 geni non presentano connessioni con la parte restante della rete: le loro interazioni geniche, che stanno alla base dei meccanismi di regolazione del livello di espressione di questo sottoinsieme di variabili, possono dunque essere considerate e analizzate separatamente da quelle che intercorrono tra i restanti 9100 geni simulati. Inoltre, poiché le alterazioni geniche simulate riguardano esclusivamente la matrice di connettività W_{900}, tra questi 900 geni rientrano tutti quelli che hanno

subito knock-out/knock-down nelle due classi di soggetti e tutte le variabili geniche che, direttamente o indirettamente regolate da tali geni, ne hanno subito l'effetto. I 185 geni definiti biomarcatori, quindi, ricadono tutti tra queste 900 variabili e la restrizione dell'analisi a questo insieme di geni non comporta la perdita di alcuna informazione circa le caratteristiche delle malattie simulate. Nel resto del capitolo si indicherà con $G = 900$ il numero di variabili a cui è stata ristretta l'analisi.

Per la selezione delle relazioni geniche significative mediante test statistico è stato adottato il criterio di correzione per test multipli basato su false discovery rate (dettagliatamente descritto nell'appendice A). In particolare il livello di significatività α è stato scelto, nella costruzione delle reti basata sia su correlazione che su correlazione parziale, in modo tale da garantire un valore della false discovery rate inferiore al 5% ($FDR \leq 5\%$).

Prima di procedere al ranking dei geni mediante le nuove misure proposte, è stata effettuata un'analisi delle proprietà delle reti di interazioni geniche costruite. L'applicazione delle misure per il ranking basate sul confronto della topologia locale delle reti A e B, infatti, presuppone che le reti siano caratterizzate da proprietà topologiche globali confrontabili. Nel resto della sezione vengono presentati alcuni risultati che riguardano il confronto delle proprietà delle reti A e B, in termini di valori di correlazione e correlazione parziale tra le coppie di geni, numero di archi, distribuzione delle connessioni tra nodi e coefficiente di clustering dei nodi.

5.1.1 Distribuzione dei valori di correlazione e correlazione parziale

Il primo passo nella costruzione delle reti di interazioni geniche consiste nel calcolo dei valori di correlazione e correlazione parziale tra tutte le possibili coppie di profili di espressione genica nelle due classi di soggetti. I boxplot delle distribuzioni dei valori ottenuti nelle classi A e B sono riportati in figura 5.1. Sia per la correlazione che per la correlazione parziale, le distribuzioni dei valori calcolati hanno forma simile nelle due classi di soggetti.

A differenza della correlazione, che assume valori nell'intervallo $[-0.9952, 0.9929]$, la correlazione parziale assume valori in un range inferiore pari a $[-0.3478, 0.3278]$. É

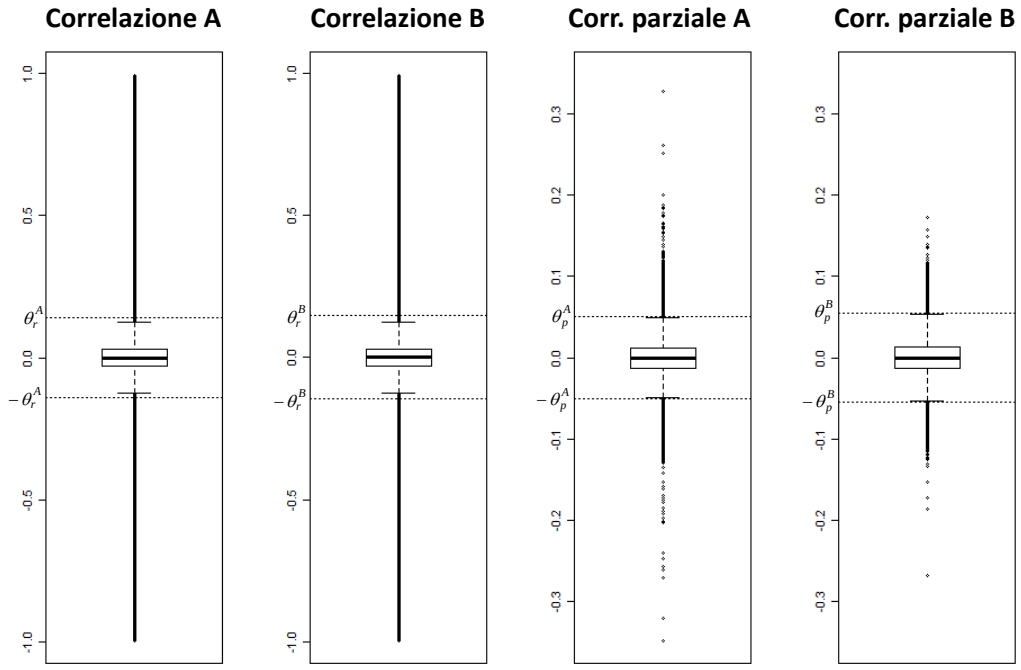


Figura 5.1 Boxplot delle distribuzioni dei valori di correlazione e correlazione parziale tra tutte le coppie di geni nelle due classi di soggetti A e B. La linea tratteggiata indica la soglia di significatività θ determinata mediante test statistico e correzione per test multipli basata su $FDR \leq 5\%$.

ragionevole attendersi una situazione di questo tipo dal momento che la correlazione parziale è una misura maggiormente restrittiva poiché tiene conto della correlazione residua tra due variabili dopo aver eseguito il condizionamento rispetto a tutte le restanti.

Una volta calcolati i valori di correlazione e correlazione parziale, a partire dalla conoscenza delle distribuzioni in ipotesi nulla (2.3) sono stati calcolati i valori della soglia di confidenza α e del rispettivo livello di significatività θ per ciascuna classe di soggetti. Tali valori, riportati nelle tabelle 5.1, sono stati determinati in modo da garantire $FDR \leq 5\%$ per le relazioni considerate significative. Anche in questo caso si osserva che α e θ assumono valori di entità confrontabile nelle due classi di soggetti, sia per la correlazione che per la correlazione parziale. Le due classi di soggetti, sebbene rappresentative di due diverse classi di malattia, sono tra loro confrontabili in termini di entità delle relazioni tra coppie di variabili geniche basate su correlazione o correlazione parziale.

correlazione		correlazione parziale	
classe A	classe B	classe A	classe B
$\alpha_r^A = 0.00151$	$\alpha_r^B = 0.00104$	$\alpha_p^A = 0.00318$	$\alpha_p^B = 0.00219$
$\theta_r^A = 0.14151$	$\theta_r^B = 0.14628$	$\theta_p^A = 0.05030$	$\theta_p^B = 0.05451$

Tabella 5.1 Valori della soglia di confidenza α e del livello di significatività θ per le due classi di soggetti con riferimento alla costruzione delle reti di relazione mediante correlazione e mediante correlazione parziale.

5.1.2 Grado di connettività

L'analisi della connettività delle reti di interazioni geniche costruite consente di ottenere informazioni circa la loro struttura topologica globale. In tabella 5.2 sono riportati i valori di alcuni parametri significativi, come ad esempio il numero di archi, il numero di nodi isolati e il massimo grado di connettività dei nodi della rete. In figura 5.2 è riportata, per le diverse reti costruite, la distribuzione del numero di connessioni che ogni nodo presenta con gli altri nodi della rete, detta degree distribution. Le reti di interazioni geniche costruite mediante entrambe le misure di similarità sono reti sparse: solo una piccola percentuale delle $E = G(G-1)/2$ possibili coppie di geni sono connesse con un arco nella rete. Gli archi presenti nelle reti costruite mediante la misura di correlazione rappresentano solo il 2-3% della totalità di possibili archi E, mentre la percentuale di archi delle reti costruite mediante correlazione parziale è circa pari a 4-5%. Si osserva inoltre una riduzione percentuale del numero di archi nelle reti costruite per la classe di soggetti B rispetto a quelli presenti nelle reti della classe A pari circa al 30%, passando da 12246 a 8400 archi per le reti costruite mediante correlazione e da 22853 a 15753 per le reti basate su correlazione parziale. Naturalmente non è detto che gli archi presenti nella rete B siano un sottoinsieme degli archi della rete A: vi possono essere archi della rete B che sono assenti nella rete A, archi della rete A assenti nella rete B, come anche archi assenti o presenti in entrambe le reti. La malattia simulata per la classe di soggetti B sembra quindi essere caratterizzata da un numero di interazioni tra variabili geniche individuate inferiore rispetto alla malattia simulata per i soggetti della classe A. Di conseguenza si osserva anche una riduzione del grado di connettività dei nodi delle reti B rispetto alle reti A: nel caso della correlazione, ad

	correlazione		correlazione parziale	
	rete A	rete B	rete A	rete B
n° archi	12246 (3%)	8400 (2%)	22853 (5,6%)	15753(3,9%)
n° nodi isolati	163	300	109	109
degree medio	27	19	51	35
degree massimo	108	91	90	59

Tabella 5.2 Parametri significativi delle reti di relazioni geniche costruite mediante correlazione e correlazione parziale: n° di archi (tra parentesi è riportata la percentuale degli archi individuati sul totale delle possibili coppie di geni), n° di nodi isolati, grado di connettività medio e massimo.

esempio, il massimo grado di connettività dei nodi della rete A, pari a 108, si riduce a 91 nella rete B. Nelle reti costruite mediante correlazione parziale il numero di nodi isolati resta invariato, mentre se ne osserva un incremento nella rete B rispetto alla rete A per quanto riguarda la ricostruzione mediante correlazione. Alcune considerazioni circa i nodi isolati, con particolare riferimento ai nodi isolati in entrambe le reti A e B, verranno riportate nella sezione 5.2.1 in cui si analizzano i risultati del ranking dei geni.

Nonostante la riduzione del grado di connettività osservata, la forma della distribuzione dei valori del grado dei nodi nelle due reti A e B non presenta differenze sostanziali, come si può osservare in figura 5.2. Le percentuali di nodi con grado di connettività molto basso, elevato o con valori intermedi sono confrontabili nelle due reti A e B per entrambe le modalità di costruzione della rete. Per quanto riguarda le reti costruite mediante correlazione, ad esempio, il grado di connettività dei nodi segue una distribuzione simile a quella di tipo power-law (4.1) sia nella rete A che nella rete B: la maggior parte dei nodi ha grado di connettività molto basso, mentre solo una piccola percentuale presenta un numero elevato di connessioni con altri nodi. La distribuzione del grado di connettività delle reti costruite mediante correlazione parziale, invece, presenta un picco in corrispondenza di valori elevati del grado di connettività: distribuzioni di questo tipo differiscono rispetto a quelle osservate per le reti di regolazione reali finora studiate. Tuttavia l'andamento simile delle distribuzioni del grado nelle reti A e B è sufficiente, anche in questo caso, a garantire la confrontabilità della struttura topologica complessiva delle reti costruite per le due classi di soggetti.

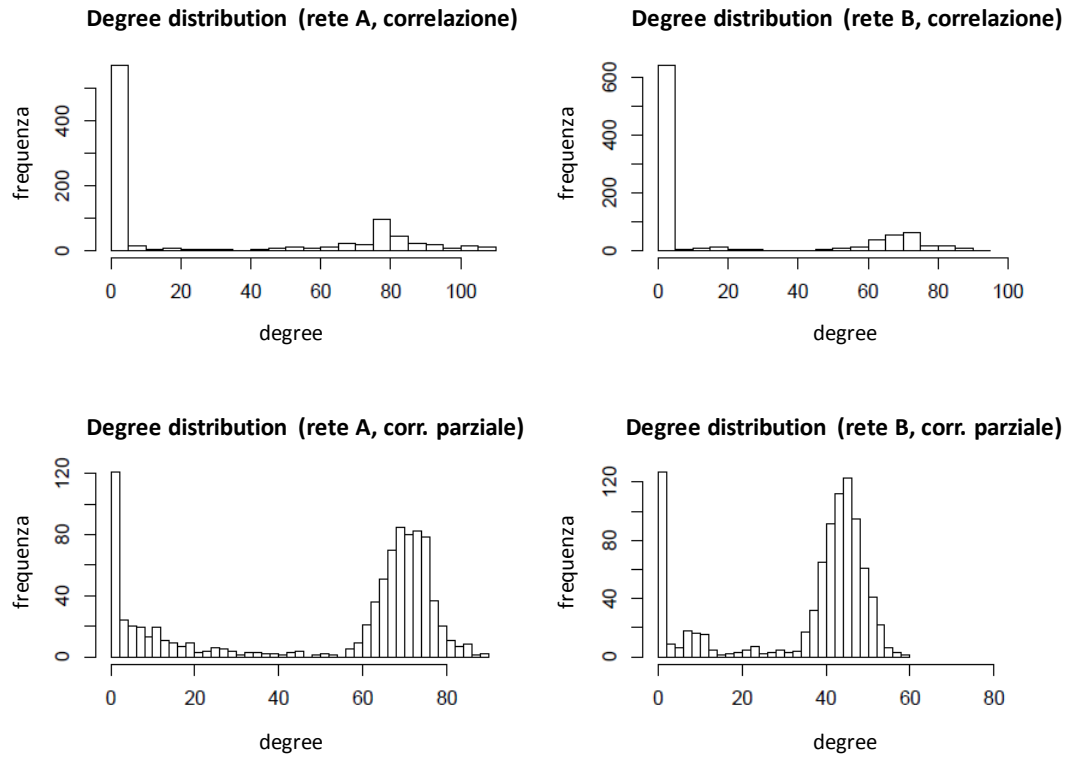


Figura 5.2 Distribuzione del grado di connettività dei nodi nelle reti A e B costruite mediante correlazione (pannelli in alto) e mediante correlazione parziale (pannelli in basso).

5.1.3 Coefficiente di clustering

Un parametro tipicamente considerato nell'analisi delle proprietà strutturali delle reti biologiche è il coefficiente di clustering medio dei nodi che quantifica la tendenza globale dei nodi della rete a formare cluster. I coefficienti di clustering medi osservati per le reti costruite riconfermano la similarità della struttura globale delle reti A e B costruite con entrambe le misure di correlazione ($CC_r^A = 0.33$ e $CC_r^B = 0.29$) e correlazione parziale ($CC_p^A = 0.14$ e $CC_p^B = 0.11$).

Un'altra grandezza presa in considerazione è il coefficiente di clustering medio dei nodi della rete con lo stesso grado di connettività. In figura 5.3 ne è riportato l'andamento osservato in funzione del grado dei nodi per le diverse reti di interazioni geniche costruite. Si osserva una buona corrispondenza tra l'andamento del coefficiente di clustering osservato per le reti A e per le reti B costruite sia mediante correlazione che mediante correlazione parziale. Anche in questo caso, quindi, è riconfermata la

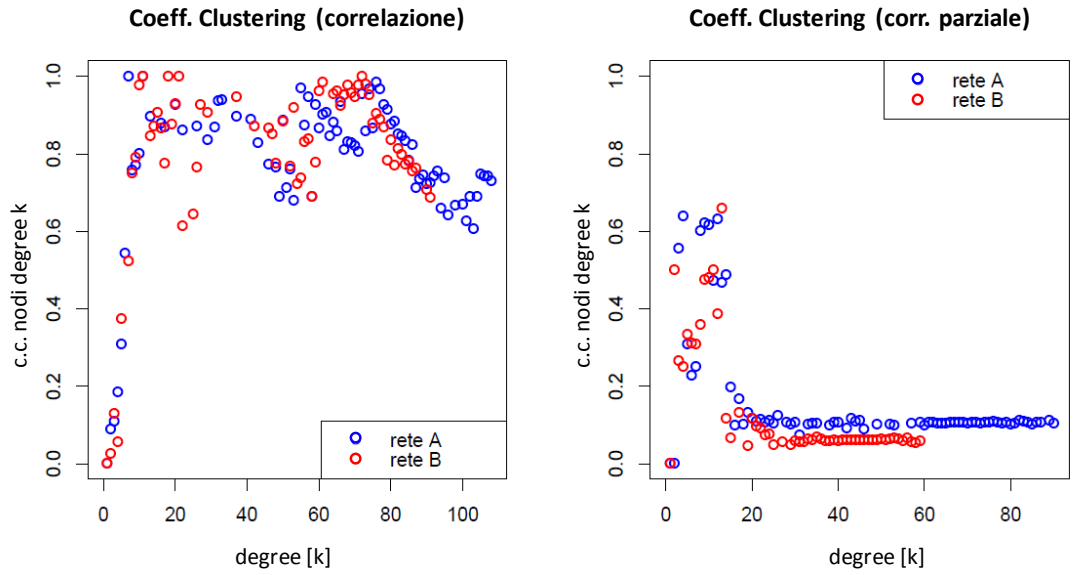


Figura 5.3 Andamento del coefficiente di clustering medio dei nodi con lo stesso grado di connettività per le reti costruite mediante correlazione (pannello a sinistra) e correlazione parziale (pannello a destra).

similarità strutturale globale delle reti ricostruite per le due classi di soggetti.

Differenze considerevoli si notano invece nell'andamento del coefficiente di clustering al variare del grado dei nodi per le due modalità di costruzione della rete. Le reti costruite mediante correlazione parziale esibiscono coefficiente di clustering decrescente all'aumentare del grado dei nodi riproducendo così l'organizzazione gerarchica strutturale che caratterizza le reti reali, in cui il coefficiente di clustering medio dei nodi con lo stesso grado varia in maniera inversamente proporzionale rispetto al grado dei nodi [72]. Nelle reti costruite mediante correlazione, al contrario, la connettività dei nodi connessi ad uno stesso nodo tende ad aumentare con il grado del nodo.

5.1.4 Signature dei nodi

Una caratterizzazione delle proprietà strutturali globali delle reti è stata fatta ricorrendo agli stessi concetti di graphlet e orbite introdotti nel capitolo 3.

Per caratterizzare complessivamente il grado di partecipazione dei nodi di una rete ai diversi graphlet, o meglio alle loro orbite, si ricorre all' estensione del concetto di degree distribution a quello di *graphlet degree distribution* [60]. Per ciascuna delle 73 orbite

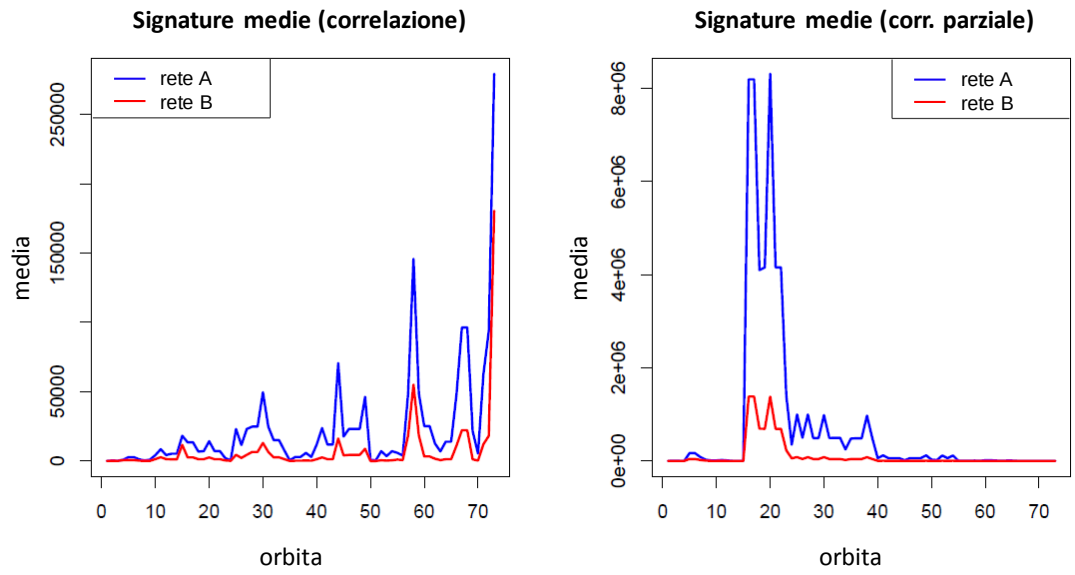


Figura 5.4 Signature medie dei nodi appartenenti alle reti A e B ottenute mediante correlazione (pannello a sinistra) e correlazione parziale (pannello a destra).

dei graphlet considerati è possibile calcolare una graphlet degree distribution (GDD) che quantifica il livello di partecipazione dei nodi della rete alla stessa orbita. In particolare la GDD i-esima misura quanti nodi partecipano all'orbita i-esima $1, 2, \ldots, k$ volte e la GDD associata all'orbita 0 corrisponde alla distribuzione del grado di connettività dei nodi. Si è osservato che le GDD calcolate per le 73 diverse orbite considerate hanno forma e andamento del tutto simili a quelli risultati per la distribuzione del grado di connettività (figura 5.2).

Parametri interessanti da considerare sono i valori medi delle GDD che definiscono quante volte in media ciascun nodo della rete compare, all'interno di un graphlet, in una delle 73 diverse orbite considerate. Il vettore dei valori medi delle GDD corrisponde alla media delle signature dei nodi di una rete. In figura 5.4 sono rappresentate le signature medie delle reti A e B costruite mediante correlazione e correlazione parziale.

Dalla figura è possibile osservare una netta differenza tra le signature medie delle reti ottenute con i due diversi metodi di costruzione, riconducibile alle diverse proprietà delle misure utilizzate per il confronto dei profili di espressione genica. Si considerino ad esempio i cinque picchi più alti delle signature medie delle reti ottenute mediante correlazione, relativi alle orbite $57, 66, 67, 71$ e 72. Nodi che appartengono a tali orbite

appartengono a graphlet con struttura caratterizzata dalla presenza di numerosi pattern di interazione "triangolari" (come si osserva in figura 3.3). Come illustrato nelle sezioni 2.3.1 e 2.3.2, la misura di correlazione non è in grado di distinguere tra interazioni di tipo diretto e indiretto: le relazioni individuate nella rete corrispondenti a interazioni indirette fanno sì che la rete costruita sia caratterizzata dalla presenza di numerosi pattern di interazione triangolari le cui diverse combinazioni danno luogo alla presenza di un numero elevato di strutture come quelle rappresentate dalle orbite $57, 66, 67, 71$ e 72. A differenza della misura di correlazione, la correlazione parziale è in grado di limitare il numero di relazioni individuate nella rete corrispondenti ad interazioni di tipo indiretto con una conseguente riduzione del numero di pattern di interazione triangolari. Come si può osservare in figura 5.4, le signature medie delle reti costruite mediante correlazione parziale assumono valori circa pari a zero in corrispondenza delle orbite $72, 57, 66, 67$ e 71. I cinque picchi più alti, invece, sono quelli relativi alle orbite $19, 15, 16, 18$ e 20 che appartengono ai graphlet con struttura tipicamente lineare.

Le signature medie delle reti A e B costruite utilizzando la stessa misura di similarità tra variabili geniche, invece, hanno un andamento simile e presentano i picchi in corrispondenza delle stesse orbite (correlazione pari a 0.927 per le signature medie delle reti costruite mediante correlazione e pari a 0.997 per quelle delle reti costruite mediante correlazione parziale). Questa corrispondenza nell'andamento delle signature medie è un'ulteriore conferma della similarità strutturale globale delle reti costruite per le due classi di soggetti A e B.

In conclusione, dal confronto delle proprietà topologiche globali delle reti costruite si può concludere che le reti ottenute, sebbene differiscano notevolmente in base alla misura di similarità adottata, sono confrontabili, per la stessa misura, per le due classi di soggetti A e B. Tale caratteristica fa sì che le differenze della topologia locale, individuate nell'intorno del nodo associato ad un gene, possano essere associate ad una effettiva variazione delle modalità di interazione e dell'attività del gene stesso e non siano piuttosto imputabili a differenze nella struttura globale delle reti messe a confronto.

L'analisi delle proprietà topologiche globali delle reti costruite ha inoltre messo in

luce alcune delle limitazioni dei metodi di ricostruzione basati su confronto pair-wise dei profili, che fanno si che tali tecniche non possano essere considerate alla stregua di vere e proprie tecniche di reverse engineering per inferire sulla rete di regolazione genica. Un esempio è dato dalla distribuzione del grado di connettività delle reti costruite mediante correlazione parziale, il cui andamento si discosta da quello di tipo power-law della rete originale. Anche l'andamento del coefficiente di clustering dei nodi, crescente con il grado di connettività dei nodi, mette in luce una delle limitazioni della misura di correlazione nel riprodurre le proprietà topologiche della rete originale. Tuttavia, come precisato nel capitolo 2, la ricostruzione della rete di regolazione genica originale esula dagli obiettivi di questo lavoro.

5.2 Ranking dei biomarcatori

La seconda fase di lavoro, dopo la costruzione delle reti di interazioni geniche, consiste nell'applicazione dei nuovi metodi proposti per il ranking dei geni. Anche in questo caso sono state considerate solamente le prime 900 variabili geniche cui è stata ristretta l'analisi nella fase di costruzione delle reti. Come già precisato nel capitolo 3, i nuovi metodi per il ranking assegnano un punteggio S_i a ciascuna variabile genica g_i che ne quantifica le variazioni di interazione con le restanti variabili nella classe di soggetti B rispetto alla classe A. Per la validazione dei nuovi metodi proposti i risultati sono stati confrontati con quelli ottenuti mediante test statistico SAM, considerato in questo contesto come metodo di riferimento e i cui dettagli sono riportati nell'appendice B. Il punteggio $S_{i,SAM}$ assegnato a ciascuna variabile genica, definito dalla (B.3), consente di ottenere delle liste ordinate le cui prime posizioni sono occupate dai geni con valori di espressione significativamente diversi nelle due classi di soggetti A e B. In figura 5.5 è proposta una rappresentazione schematica delle fasi di lavoro e dei diversi score calcolati.

Ordinando le variabili geniche secondo un ordinamento decrescente dei punteggi ad esse associati è possibile ottenere un ranking di geni in cui i primi posti sono occupati dai geni per i quali è stata rilevata la massima variazione delle interazioni con le restanti variabili. Lavorando con dati simulati e conoscendo a priori la lista dei 185 geni

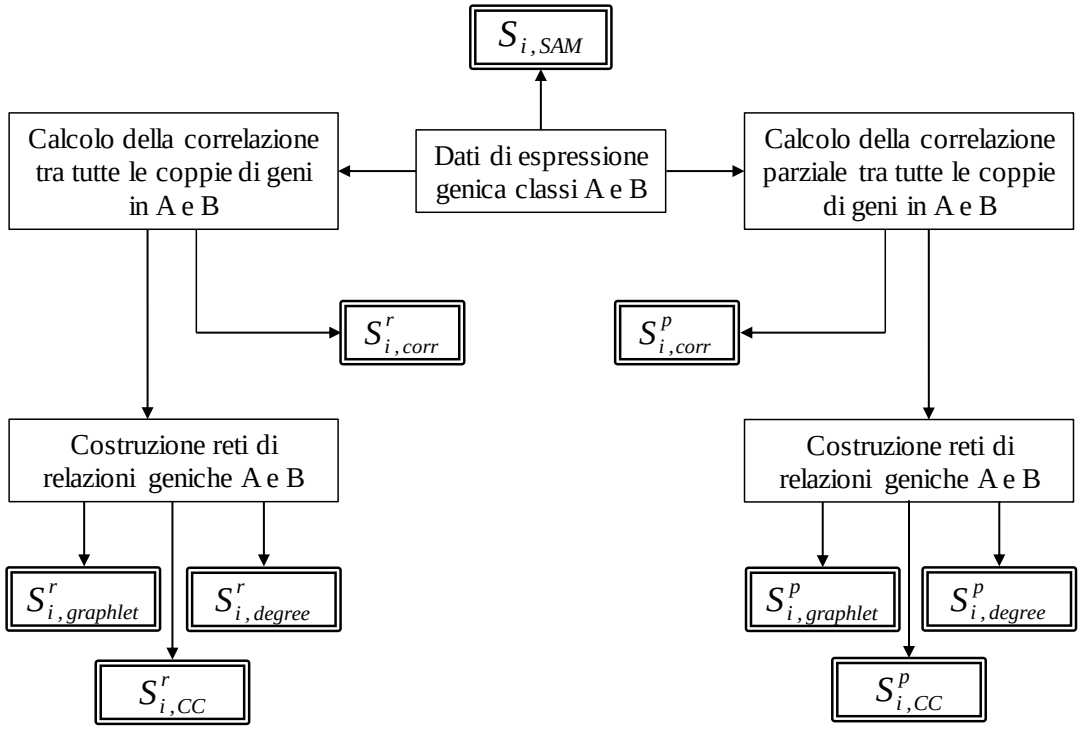


Figura 5.5 Rappresentazione schematica dei diversi score ottenuti: gli apici r e p indicano che gli score sono stati ottenuti utilizzando rispettivamente la correlazione o la correlazione parziale come misura di similarità.

biomarcatori è possibile valutare la capacità dei metodi di assegnare punteggi elevati ai geni biomarcatori. L'obiettivo è quello di verificare la validità del nuovo criterio proposto per il ranking dei geni valutando la robustezza delle misure adottate per la rilevazione delle variazioni di interazioni geniche significative e per l'identificazione dei geni biomarcatori. Per questa prima valutazione delle prestazioni le diverse liste ordinate sono state messe a confronto e sono state confrontate le posizioni nel ranking dei geni biomarcatori.

Si consideri una lista di geni ordinata secondo un ordinamento decrescente degli score ad essi associati e si supponga di selezionare i primi k geni con punteggio maggiore. Si definisce *precisione* il rapporto tra il numero di biomarcatori selezionati BM_{sel}, ovvero il numero di geni tra i k selezionati che compaiono nella lista dei

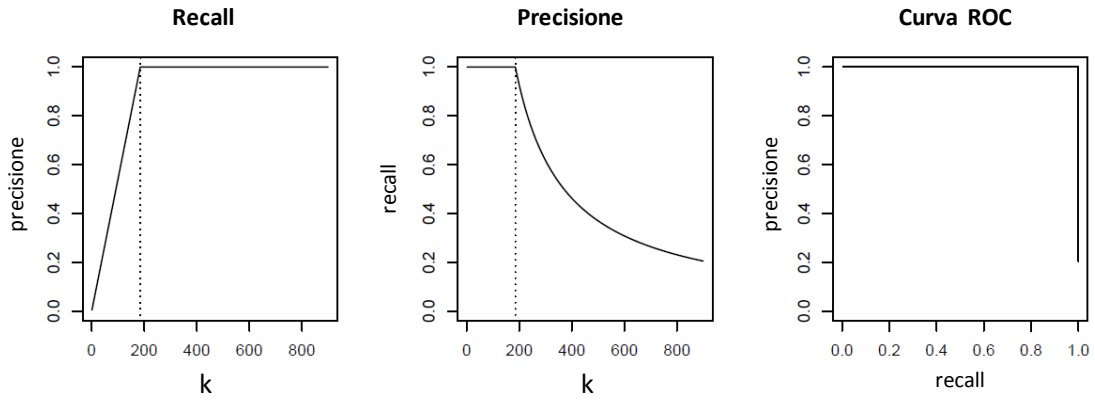


Figura 5.6 Andamento di precisione e recall al variare di k e curva ROC per una situazione di "ranking ideale" dei $G = 900$ geni in cui le prime posizioni della lista ordinata sono occupate dai $BM = 185$ biomarcatori. La linea verticale tratteggiata in corrispondenza di $k = BM$ indica il valore di k per il quale la recall assume valore unitario e la precisione inizia a decrescere.

biomarcatori, e il numero totale k di geni selezionati:

$$\textit{Precisione}_k = \frac{BM_{sel}}{k}.$$

Si definisce invece *recall* il rapporto tra il numero di biomarcatori selezionati e il numero totale di biomarcatori BM:

$$\textit{Recall}_k = \frac{BM_{sel}}{BM}.$$

Sia la precisione che la recall assumono valori nel range $[0, 1]$.

Calcolando i valori di precisione e recall al variare del numero k di geni ordinati secondo un determinato score è possibile valutare la bontà del ranking dei geni ottenuto. Uno score ideale assegna i punteggi più elevati ai BM geni biomarcatori che occupano quindi le prime posizioni della lista ordinata di geni, mentre i restanti $G - BM$ geni occupano le ultime posizioni della lista. In questo caso ideale il valore di precisione è massimo e pari a 1 per tutti i valori di k che vanno da 1 a BM (come si osserva in figura 5.6, pannello centrale) e assume valori via via decrescenti per i $k > BM$, per i quali tra i k geni selezionati viene incluso un numero via via crescente di geni non biomarcatori. La recall, invece, che definisce la percentuale di biomarcatori selezionati sul totale dei biomarcatori, assume valori crescenti per $k = 1, \ldots, BM - 1$, mentre assume valore unitario per i $k \geq BM$ in corrispondenza dei quali i geni biomarcatori, che occupano

le prime BM posizioni della lista, sono stati tutti selezionati (figura 5.6, pannello a sinistra). Una rappresentazione compatta di entrambe le misure di precisione e recall è data dalla curva ROC (*Receiver Operating Characteristic*) che presenta sull'asse delle ascisse i valori di recall e sull'asse delle ordinate i valori di precisione. Nella situazione ideale precedentemente descritta la curva ROC è caratterizzata da un primo tratto in cui la precisione assume valore costantemente pari a 1 in corrispondenza di valori crescenti di recall; solo quando la recall assume valore unitario, ad indicare che tutti i geni biomarcatori sono stati selezionati, la precisione decresce fino a raggiungere il valore finale di BM/G (figura 5.6, pannello a destra). Il calcolo dell'area sottesa dalla curva ROC fornisce un parametro il cui valore caratterizza le prestazioni del metodo per il ranking dei geni: l'*Area Under Curve* (AUC) assume valore unitario nella situazione di ranking ideale.

5.2.1 Analisi qualitativa degli score

Prima di analizzare le liste ordinate di geni ottenute con i diversi metodi di ranking proposti, può essere utile fare alcune considerazioni circa i valori assunti dai diversi score calcolati per i 900 geni presi in esame. Le osservazioni riportate in questa sezione fanno riferimento ai grafici di figura 5.7 e 5.8. In figura 5.7 sono riportati i valori dei diversi score ottenuti utilizzando la correlazione come misura di similarità tra geni, mentre in figura 5.8 quelli ottenuti utilizzando la correlazione parziale. I grafici nella parte superiore delle figure rappresentano (in nero) i valori degli score calcolati per i geni non biomarcatori, mentre nei grafici della parte inferiore sono raffigurati (in verde) gli score calcolati per i geni biomarcatori. Tra i 185 biomarcatori è possibile distinguere i 6 geni G^B_{KOKD} sui quali è stato effettuato direttamente il knock-out/knock-down per simulare la malattia della classe di soggetti B, i cui score sono rappresentati in rosso. Nella prima colonna di entrambe le figure sono riportati come confronto gli score calcolati mediante test SAM, mentre nelle restanti colonne sono riportati gli score calcolati ricorrendo alle quattro diverse misure definite.

Tutti gli score utilizzati per il ranking dei geni sono score normalizzati che possono assumere valori esclusivamente nell'intervallo $[0, 1]$. La prima cosa che si può osservare

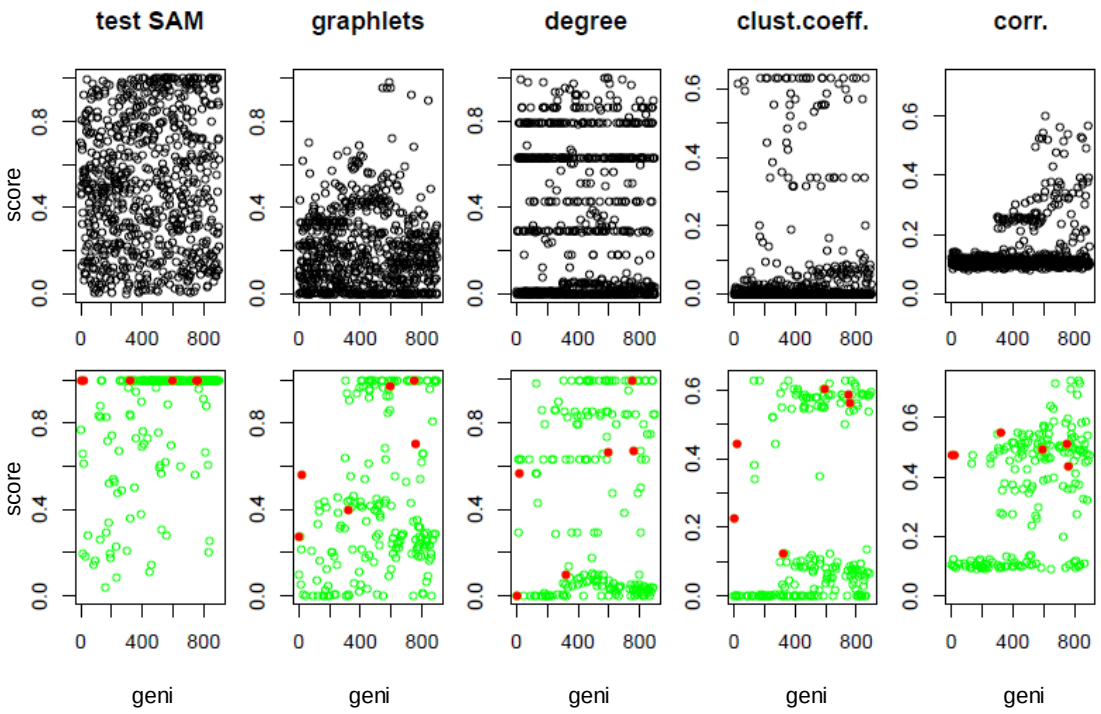


Figura 5.7 Valori dei diversi score ottenuti utilizzando la correlazione come misura di similarità. Nei pannelli superiori sono riportati gli score dei geni non biomarcatori, in quelli inferiori gli score dei biomarcatori.

è che lo score che quantifica la massima variazione di correlazione parziale tra coppie di geni (ultima colonna in figura 5.8), a differenza degli altri, assume valori in un intervallo molto più ristretto: il massimo valore osservato è infatti pari a 0.175. Questo dipende in parte dal fatto che i valori stessi di correlazione parziale tra tutte le possibili coppie di geni assumono valori in un intervallo ridotto pari circa a $[-0.35, 0.35]$ (come si osserva in figura 5.2), ma è soprattutto indicativo del fatto che non si osservano variazioni significative della correlazione parziale tra coppie di geni dalla classe A alla classe B.

L'osservazione dei grafici degli score mette inoltre in luce la capacità della misura basata su test SAM di assegnare punteggi elevati, prossimi a 1, alla maggior parte dei geni biomarcatori: 122 biomarcatori su 185 sono caratterizzati da uno score superiore a 0.98. A differenza degli altri score, quello basato su test SAM è l'unico che attribuisce punteggio massimo a tutti i 6 geni G^B_{KOKD} che hanno subito direttamente knock-out/knock-down. Tuttavia l'assegnazione di un punteggio prossimo a 1 anche

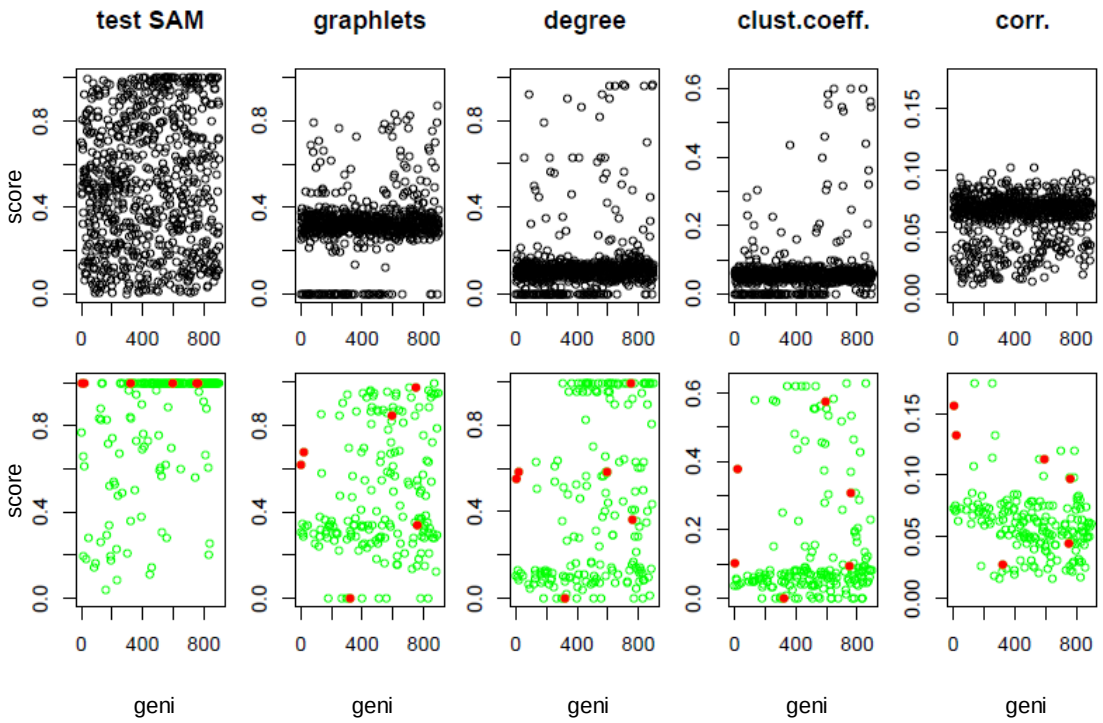


Figura 5.8 Valori dei diversi score ottenuti utilizzando la correlazione parziale come misura di similarità. Nei pannelli superiori sono riportati gli score dei geni non biomarcatori, in quelli inferiori gli score dei biomarcatori.

ad un numero elevato di geni non biomarcatori riduce le prestazioni del metodo: i 122 biomarcatori costituiscono solo il 70% dei geni con score superiore a 0.98.

A differenza dei metodi basati su test SAM, quelli basati su graphlet, applicati sia alle reti costruite mediante correlazione che mediante correlazione parziale, assegnano score bassi nella scala da 0 a 1 ad un numero elevato di geni biomarcatori. Tuttavia, mentre nel caso dei metodi basati su test SAM i geni non biomarcatori sono caratterizzati da score che assumono valori, secondo una distribuzione uniforme, in tutto il range $[0, 1]$, nei metodi basati su graphlet solo gli score di alcuni geni biomarcatori assumono valori elevati permettendo di differenziarli dal resto della distribuzione che assume valori inferiori. A differenza del metodo basato su test SAM è dunque possibile individuare un valore di soglia al di sopra del quale ricadano gli score esclusivamente di geni biomarcatori, o comunque di un numero estremamente ridotto di geni non biomarcatori. Si supponga ad esempio, nel caso delle reti costruite mediante correlazione, di fissare

un valore di soglia pari a 0.73: i geni il cui score supera tale valore sono in tutto 53, 47 dei quali sono biomarcatori (ovvero l'88%). Nel caso delle reti costruite mediante correlazione parziale, invece, è possibile individuare un valore di soglia pari a 0.88 in modo tale che i 25 geni con score superiore a 0.88 siano tutti biomarcatori.

Già da questa prima analisi di tipo qualitativo si intuisce che il metodo basato sulla quantificazione della variazione del coefficiente di clustering non è caratterizzato da buone prestazioni. Gli score basati su coefficiente di clustering, ottenuti dal confronto delle reti A e B costruite sia mediante correlazione che mediante correlazione parziale, assumono infatti valori simili per i geni biomarcatori e non biomarcatori rendendo particolarmente difficile l'identificazione dei biomarcatori unicamente sulla base di questo punteggio.

Osservando i grafici in figura 5.7 e 5.8 si può constatare che gli unici metodi per i quali sono stati calcolati score con valore esattamente pari a zero sono quelli basati sul confronto della topologia locale delle reti di interazioni geniche. Score nulli sono assegnati non solo ad un certo numero di geni non biomarcatori ma anche ad una percentuale di geni appartenenti alla lista dei biomarcatori che vengono quindi collocati, contrariamente a quanto desiderato, nelle ultime posizioni delle liste di geni ordinate. L'assegnazione di score nulli a geni biomarcatori, che si verifica utilizzando sia la correlazione che la correlazione parziale come misure di similarità, può essere in parte dovuta alla presenza di geni biomarcatori associati a nodi che sono isolati in entrambe le reti da confrontare. La costruzione delle reti basata su correlazione determina la presenza di 10 nodi associati a biomarcatori e isolati in entrambe le reti A e B, mentre la costruzione delle reti mediante correlazione parziale ne genera 9, uno dei quali appartiene all'insieme dei 6 geni G^B_{KOKD}. Mentre è ragionevole attendersi che alcuni geni biomarcatori non presentino interazioni con altri geni nella classe di soggetti B dove sono stati sottoposti a knock-out/knock-down (o ne subiscono gli effetti), la presenza di biomarcatori isolati nella rete A è insolita poiché i geni della malattia sono stati scelti proprio tra i geni che presentano un elevato numero di connessioni con altri geni. L'insolita presenza di geni biomarcatori isolati in entrambe le reti A e B è riconducibile a intrinseche limitazioni delle tecniche utilizzate per la costruzione

delle reti di interazioni geniche. Il metodo basato sul confronto pair-wise dei profili di espressione genica, infatti, consente di tenere in considerazione soltanto le relazioni tra coppie di variabili trascurando invece pattern di interazione più complessi.

Per quanto riguarda i metodi basati sulla quantificazione della massima variazione di correlazione si può osservare che, utilizzando come misura di similarità la correlazione parziale, non si ottengono buone prestazioni. Infatti, sebbene vi siano une decina di biomarcatori i cui score assumono valori superiori al valore massimo assunto dagli score dei geni non biomarcatori, la maggior parte dei biomarcatori è caratterizzata da score con valori nello stesso range di quelli dei restanti geni. Utilizzando come misura di similarità la correlazione, invece, il metodo sembra avere prestazioni decisamente migliori: gli score calcolati per la maggior parte dei biomarcatori assumono valori che si discostano significativamente dal valor medio degli score associati ai restanti geni.

5.2.2 Curve ROC: precisione e recall

In figura 5.9 sono rappresentate le curve ROC ottenute per i vari metodi considerati, mentre in tabella 5.3 sono riportati i relativi valori di AUC.

Una prima osservazione delle curve ROC ottenute conferma quanto già si intuiva dall'analisi qualitativa degli score, ovvero che le prestazioni del test SAM restano complessivamente superiori a quelle dei nuovi metodi di ranking proposti. La curva ROC del test SAM (rappresentata in nero) è infatti quella che più si avvicina a quella ideale e il corrispondente valore di AUC è il maggiore osservato ($AUC_{SAM} = 0.71$). Tuttavia alcuni dei nuovi metodi di ranking, in particolare quello basato su graphlet, consentono di ottenere delle liste di geni ordinate caratterizzate da precisione iniziale pari a 1, a differenza del test SAM caratterizzato da precisione iniziale inferiore. Questo dipende dal fatto che il test SAM assegna punteggio massimo (corrispondente a $p\text{-}value = 0$) a un numero elevato di geni (114) tra i quali compaiono numerosi biomarcatori (103) ma anche alcuni geni non biomarcatori (11).

Tra i nuovi metodi di ranking proposti quello che presenta le prestazioni migliori, che più si avvicinano a quelle del test SAM, è il metodo basato sulla massima variazione di correlazione tra coppie di geni. Tale metodo è caratterizzato da un valore di AUC pari

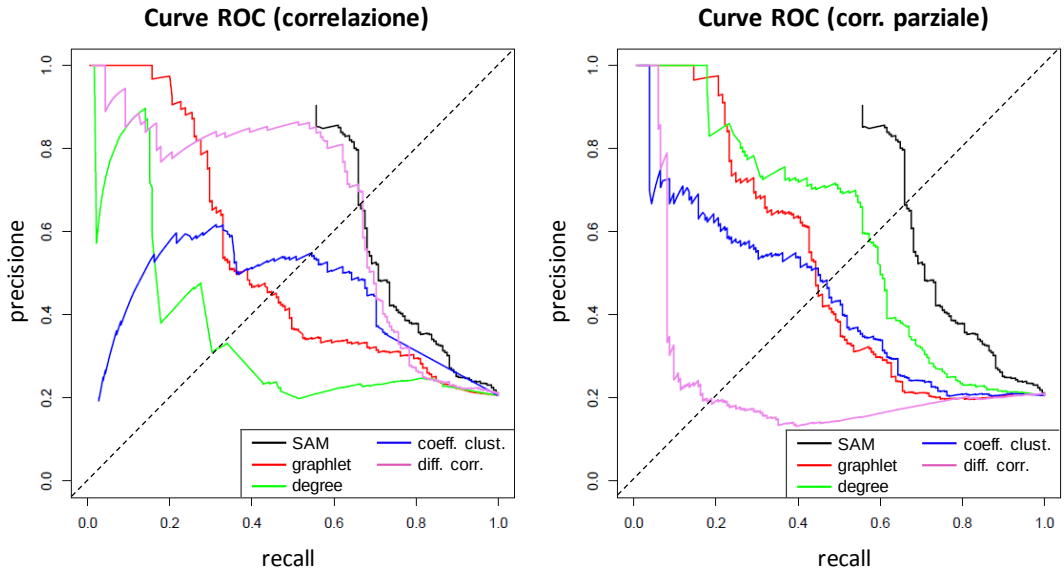


Figura 5.9 Curve ROC ottenute per i diversi metodi di ranking dei geni confrontate con quella ottenuta mediante test SAM. Nel grafico di sinistra sono riportate le curve ROC ottenute utilizzando la correlazione come misura di similarità, nel grafico di destra quelle ottenute utilizzando la correlazione parziale. La bisettrice individua i punti delle curve ROC in corrispondenza dei quali il numero di geni selezionati è pari al numero di biomarcatori e che, in un ranking ideale, dovrebbero essere caratterizzati da precisione e recall entrambe pari a 1.

a 0.66, superiore a quello di tutti gli altri metodi e che si avvicina a quello calcolato per il test SAM. La curva ROC (rappresentata in rosa nel grafico di sinistra) è caratterizzata da valore di precisione iniziale pari a 1, la precisione si mantiene quasi sempre al di sopra del valore 0.8 fino a raggiungere valori elevati di recall: la precisione decresce rapidamente soltanto per valori di recall superiori a 0.6, esattamente come per il test SAM. Questo significa che la percentuale di geni biomarcatori selezionati, sul totale dei geni selezionati, resta sempre superiore all'80% finchè non viene selezionato il 60% dei geni biomarcatori.

Lo stesso metodo applicato utilizzando però come misura di similarità la correlazione parziale, è quello che esibisce le peggiori prestazioni in termini di precisione e recall. La precisione, inizialmente pari a 1, decresce rapidamente fino a raggiungere valori inferiori a 0.2 in corrispondenza di valori di recall molto bassi (minori di 0.2). L'Area Under Curve è pari a 0.24 e corrisponde al minimo valore osservato per il parametro AUC.

Questa sostanziale differenza delle prestazioni dello stesso metodo applicato adottando la correlazione o la correlazione parziale è riconducibile alle diverse caratteristiche e proprietà delle due misure di similarità. La misura di correlazione semplice tra due variabili geniche ne quantifica non soltanto la relazione diretta, ma anche gli eventuali contributi legati alla presenza di variabili esterne alla coppia e che influenzano l'attività di entrambi i geni. Al contrario, la misura di correlazione parziale, mediante il condizionamento rispetto a tutte le restanti variabili, è in grado di quantificare la relazione diretta esistente tra i geni eliminando il contributo legato ad altre variabili. Per questa sua caratteristica la correlazione parziale è una misura più accurata ed affidabile per l'identificazione di relazioni dirette tra geni. Tuttavia, mentre la correlazione parziale quantifica unicamente le variazioni dell'interazione diretta tra due geni, la correlazione semplice consente di rilevare anche quelle dovute a una variazione delle interazioni delle due variabili con i restanti geni. In questo contesto la misura di correlazione risulta essere maggiormente informativa poiché è in grado di dare una caratterizzazione più completa della variazione delle modalità di interazione tra geni nelle due classi di soggetti considerate.

Un altro motivo che determina la riduzione delle prestazioni osservata nei grafici può essere dovuto al fatto che la misura di correlazione parziale è maggiormente sensibile al rumore e alla variabilità biologica che caratterizzano i dati. Poiché la correlazione parziale misura soltanto una frazione della correlazione tra due geni, ovvero quella residua dopo il condizionamento rispetto a tutte le restanti variabili, si può ipotizzare che questa risenta maggiormente del rumore associato ai dati rispetto alla misura di correlazione semplice. La scarsa attendibilità e consistenza dei valori di correlazione parziale calcolati si ripercuote direttamente sulle prestazioni del metodo di ranking basato sulla variazione di correlazione parziale tra coppie di geni.

Una riduzione delle prestazioni utilizzando la correlazione parziale piuttosto che quella semplice non si osserva invece per i metodi di ranking basati sul confronto delle proprietà topologiche delle reti di interazioni geniche. Si può supporre che sia proprio l'introduzione di una soglia di significatività a limitare l'effetto del rumore sui valori di correlazione parziale. Durante il procedimento di costruzione delle reti, infatti,

	Area Under Curve (AUC)	
	correlazione	corr. parziale
test SAM	0.71	0.71
graphlet	0.53	0.51
degree	0.36	0.60
coeff. di clustering	0.44	0.44
correlazione	0.66	0.24

Tabella 5.3 Valori dell'Area Under Curve (AUC) per i diversi metodi di ranking e per le due diverse misure di similarità adottate confrontati con quello ottenuto mediante test SAM.

vengono prese in considerazione soltanto le relazioni geniche significative caratterizzate da valore di correlazione parziale superiore ad una soglia di significatività. In questo modo si escludono dall'analisi quelle variazioni di interazione che hanno maggiormente risentito della presenza di rumore.

Tra i metodi basati su confronto della topologia delle reti di interazioni geniche, quello basato su coefficiente di clustering esibisce le prestazioni peggiori in termini di precisione e recall. La lista ordinata di geni ottenuta applicandolo alle reti costruite mediante correlazione è caratterizzata da una curva ROC con valori di precisione inizialmente molto bassi e che restano sempre inferiori a 0.6. L'applicazione alle reti costruite mediante correlazione parziale consente di ottenere valori unitari di precisione per le prime posizioni della lista, tuttavia la precisione diminuisce velocemente per bassi valori di recall e l'Area Under Curve misurata, pari a 0.44, è una delle più basse. Alla luce di questi risultati si può ipotizzare che l'osservazione delle variazioni di interazione tra i geni direttamente connessi ad uno stesso gene nelle due reti non è adeguata per l'identificazione dei biomarcatori.

Risultati migliori si ottengono utilizzando i metodi basati sul confronto del grado di connettività dei nodi o della partecipazione ai graphlet e alle loro orbite. Il metodo basato sul confronto del grado di connettività dei nodi se applicato alle reti costruite mediante correlazione consente di ottenere una lista ordinata di geni caratterizzata da valori di precisione che variano con discontinuità evidenziando la scarsa abilità del metodo nel distinguere i geni biomarcatori dai restanti geni. Il metodo dà invece risultati migliori quando applicato alle reti costruite mediante correlazione parziale. Questo

miglioramento delle prestazioni può essere dovuto al fatto che una misura semplice come quella basata esclusivamente sul confronto del numero di interazioni dei geni risente in modo particolare della presenza dell'elevato numero di relazioni di tipo falso positivo individuate mediante la misura di correlazione. Ricorrendo alla correlazione parziale si limitano le relazioni di tipo falso positivo individuate e anche una misura semplice come quella basata sul grado riesce ad individuare con una precisione superiore le variazioni di interazione significative a carico dei geni biomarcatori.

Il metodo basato su graphlet è invece caratterizzato da una maggior robustezza relativamente alla presenza delle relazioni di tipo falso positivo individuate mediante correlazione poiché consente di ottenere delle buone prestazioni sia se applicato alle reti ottenute mediante correlazione parziale che mediante correlazione semplice. Il metodo basato su graphlet, inoltre, supera tutti gli altri metodi in termini di precisione iniziale delle liste. Confrontando la partecipazione dei nodi ai vari graphlet considerati e alle loro orbite vengono infatti assegnati score massimi ad un numero relativamente elevato di geni biomarcatori: dal confronto delle reti costruite sia mediante correlazione che mediante correlazione parziale si ottengono due liste ordinate di geni le cui prime 29 e 27 posizioni, rispettivamente, sono occupate unicamente da geni biomarcatori. Di questi biomarcatori, 20 compaiono nelle prime posizioni di entrambe le liste ottenute: viene dunque riconfermata la robustezza del metodo che, indipendentemente dalla modalità di costruzione della rete, riesce ad individuare gli stessi 20 biomarcatori assegnando loro i punteggi massimi. Tuttavia, come evidenziato dalle curve ROC, il metodo basato su graphlet è caratterizzato da una diminuzione della precisione in corrispondenza di valori relativamente bassi di recall (inferiori a 0.4).

Una considerazione interessante riguarda il fatto che, se applicato alle reti costruite mediante correlazione, il metodo basato su graphlet esibisce prestazioni decisamente superiori rispetto a quelle ottenute mediante confronto del grado di connettività dei nodi. Si osservino infatti l'andamento delle curve ROC e i relativi valori di AUC: la curva ROC del metodo basato su graphlet si mantiene sempre al di sopra di quella del metodo basato sul grado dei nodi (grafico di sinistra in figura 5.9) ed è caratterizzata da valore di AUC superiore (0.53 contro 0.36 per il metodo basato sul grado). Sulla base di questo

risultato si potrebbe concludere che una misura come quella basata su graphlet consente di ottenere una miglior caratterizzazione delle modalità di interazione genica e di conseguenza una più robusta e consistente quantificazione della loro variazione rispetto a misure più semplici come quella basata sul grado o sul coefficiente di clustering. Tuttavia la capacità del metodo basato su graphlet di rilevare le variazioni di interazione significative a carico dei geni biomarcatori può essere messa in discussione osservando i risultati ottenuti per le reti costruite mediante correlazione parziale. In questo caso infatti una misura molto più semplice basata esclusivamente sul confronto del grado di connettività dei nodi consente di ottenere risultati migliori rendendo discutibile l'utilità di ricorrere ad una caratterizzazione complessa delle interazioni geniche come quella fornita dai graphlet.

In conclusione si può dire che i metodi basati sul confronto della topologia locale delle reti non consentono di ottenere prestazioni migliori di quello basato sulla variazione di correlazione semplice. I primi, infatti, si basano su una costruzione delle reti di interazioni geniche che tiene conto unicamente del modulo dei valori di correlazione calcolati perdendo quindi l'informazione sul segno. Inoltre, l'introduzione di una soglia di significatività, porta ad una sorta di discretizzazione secondo la quale la relazione tra due geni viene considerata significativa o non significativa perdendo l'informazione sul peso della correlazione. Il metodo basato sulla variazione di correlazione al contrario, consente di tenere in considerazione sia il segno che il peso associato ai valori di correlazione. É ragionevole ipotizzare che sia proprio questo il valore aggiunto del metodo che consente di individuare con precisione superiore le variazioni di interazione significative legate proprio ai geni biomarcatori. Alla luce dei risultati, infatti, il metodo basato sulla variazione di correlazione consente di tenere in considerazione delle informazioni che non vengono invece rilevate dai metodi basati sul confronto delle reti. Si osserva, ad esempio, che circa il 90% delle massime variazioni di correlazione rilevate si riferiscono proprio a casi in cui è avvenuto un cambio di segno della correlazione nelle due classi di soggetti. Inoltre, nel 60% di questi casi la variazione di correlazione non viene rilevata mediante confronto delle reti poiché l'introduzione della soglia di significatività porta alla presenza o assenza del relativo arco nelle reti di entrambe le classi di soggetti.

5.3 Riduzione del numero di campioni

I risultati finora proposti sono relativi all'applicazione dei nuovi metodi per l'identificazione dei biomarcatori alle due classi di popolazione simulate A e B, ciascuna costituita da 500 soggetti. Un numero di campioni così elevato è raramente disponibile nella pratica clinica: sia per motivi tecnici ed etici che per ragioni di costo il numero di campioni tipicamente ottenuti mediante esperimenti con microarray è di qualche decina. Per questo motivo è particolarmente interessante analizzare come variano le prestazioni dei nuovi metodi di ranking applicati a dataset costituiti da un numero ridotto di campioni per ciascuna delle variabili geniche monitorate.

A questo scopo il dataset di partenza, costituito dalle due classi A e B di 500 soggetti ciascuna, è stato partizionato in modo tale da ottenere 10 diversi dataset ciascuno costituito da due classi di 50 soggetti. In particolare le classi di partenza A e B sono state frazionate ciascuna in 10 classi da 50 soggetti: la suddivisione dei soggetti è stata operata in modo tale da garantire, nelle nuove classi ottenute, la stessa eterogeneità riprodotta per le malattie A e B nelle due classi di soggetti di partenza. I dieci dataset così ottenuti, costituiti da un numero decisamente inferiore di campioni, sono rappresentativi delle stesse due classi di popolazione iniziali, di cui riproducono le caratteristiche.

L'applicazione dei nuovi metodi di ranking e del metodo di riferimento basato su test SAM è stata dunque riproposta per i 10 dataset così ottenuti. Per ciascun dataset sono state costruite le reti di interazioni geniche A e B relative alle due classi di soggetti utilizzando come misure di similarità la correlazione e la correlazione parziale. Dal confronto delle reti costruite, o semplicemente dal confronto dei valori di correlazione ottenuti, sono stati poi calcolati gli score associati a ciascuna variabile genica utilizzando le diverse misure proposte. Per ciascuno dei metodi di ranking sono dunque state ottenute 10 liste di geni ordinate in base allo score, ciascuna delle quali relativa al confronto delle due classi A e B di uno dei 10 dataset.

L'applicazione dei metodi di ranking ai dieci diversi dataset, rappresentativi dello stesso dataset iniziale, consente di fare delle considerazioni anche circa la riproducibilità dei risultati ottenuti. Una delle limitazioni delle tecniche di feature selection comune-

mente utilizzate è infatti dovuta alla variabilità delle liste di biomarcatori individuate e, più in generale, alla variabilità del ranking dei geni determinato. La ricerca dei geni biomarcatori mediante analisi di dati di microarray condotta nei laboratori e nei centri di ricerca dà spesso luogo a risultati differenti, mettendo così in discussione l'affidabilità e la robustezza delle liste di geni individuate. Questa variabilità dei risultati è legata alla disponibilità di un numero di campioni limitato rispetto all'elevato numero di variabili monitorate [6], alla scarsa riproducibilità dei protocolli sperimentali adottati dai diversi laboratori [8, 9], ma anche alla intrinseca complessità ed eterogeneità delle patologie analizzate caratterizzate da mutazioni a carico di diversi geni e dall'alterazione di pathway di regolazione differenti [10].

La simulazione della variabilità biologica dei soggetti e dell'eterogeneità della malattia delle due classi di popolazione e la suddivisione dei 500 soggetti nei 10 dataset consente di valutare la riproducibilità dei risultati ottenuti mediante i diversi metodi di ranking in presenza di un numero limitato di campioni. Per confrontare tra loro le liste di geni ordinate ottenute con lo stesso metodo di ranking si ricorre al concetto di "stabilità" delle liste, che verrà definito nella sezione 5.3.2.

5.3.1 Curve ROC: precisione e recall

Le liste ordinate ottenute con i diversi score e per i diversi dataset possono essere confrontate ricorrendo alle stesse misure di precisione e recall precedentemente definite che danno informazioni circa le posizioni occupate dai geni biomarcatori nelle liste. In particolare dall'analisi delle curve ROC medie e dal confronto dei valori medi dell'Area Under Curve, determinati mediando i valori di precisione e recall ottenuti nei dieci dataset per ciascuno degli score, è possibile valutare se e come variano le prestazioni dei diversi metodi di ranking in corrispondenza della riduzione del numero di campioni da 500 a 50. In figura 5.10 sono riportate le curve ROC medie relative ai diversi metodi di ranking.

Come ci si attende, in corrispondenza ad una riduzione del numero di campioni disponibili da 500 a 50 si osserva un peggioramento, più o meno evidente, delle prestazioni di tutti i metodi di ranking. Si confronti, in particolare, la curva ROC media

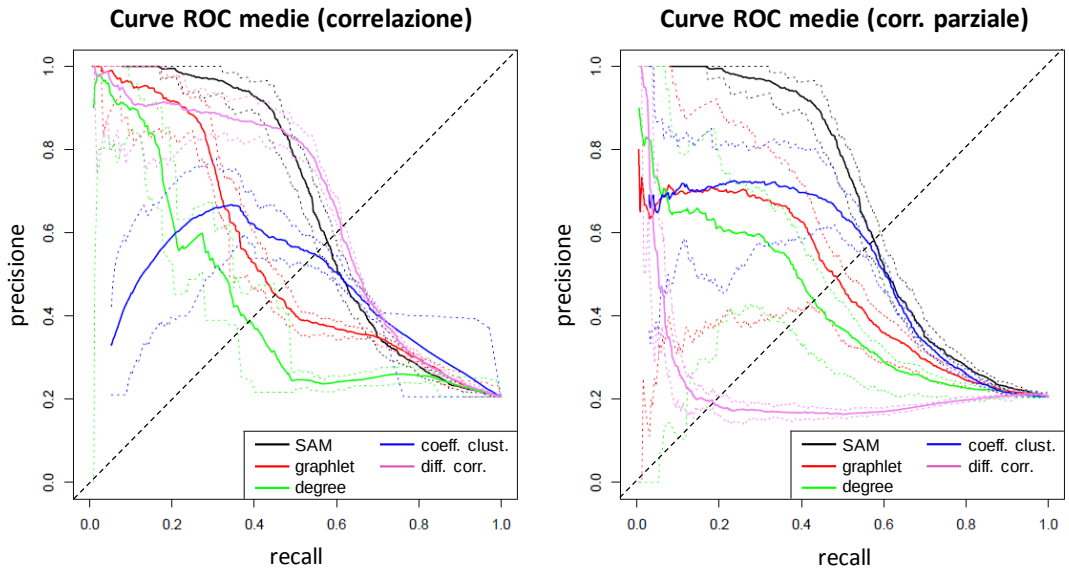


Figura 5.10 Curve ROC medie ottenute applicando i diversi metodi di ranking ai 10 dataset da 50 soggetti ciascuno. Le linee tratteggiate indicano il massimo e il minimo valore di precisione ottenuti. Nel grafico di sinistra sono riportate le curve ROC ottenute utilizzando la correlazione come misura di similarità, nel grafico di destra quelle ottenute utilizzando la correlazione parziale.

relativa all'applicazione del test SAM sui 10 dataset rappresentata in nero in figura 5.10 con quella riportata in figura 5.9: nonostante l'elevata precisione iniziale, la precisione decresce in corrispondenza di valori inferiori di recall determinando una diminuzione dell'Area Under Curve media da 0.71 a 0.67.

I metodi le cui prestazioni, mediamente, risentono meno della riduzione del numero di campioni sono quelli che ricorrono alla misura di correlazione semplice e che si basano sia sul confronto della topologia delle reti di interazioni geniche che sul confronto diretto dei valori di correlazione tra coppie di geni. Questo mette in evidenza la robustezza della misura di correlazione semplice che, anche in presenza di un numero limitato di campioni, è in grado di produrre una quantificazione attendibile delle variazioni di interazione rilevando quelle più significative a carico dei geni biomarcatori.

In particolare, riducendo il numero di campioni a disposizione, le prestazioni del metodo basato su confronto dei valori di correlazione sono simili a quelle ottenute con il test SAM, sottolineando la validità del metodo nell'individuare le relazioni

	sd media precisione/recall	
	correlazione	corr. parziale
test SAM	0.018/0.029	0.018/0.029
graphlet	0.014/0.016	0.043/0.040
degree	0.020/0.024	0.042/0.046
coeff. di clustering	0.016/0.012	0.023/0.023
correlazione	0.013/00.017	0.016/0.018

Tabella 5.4 Valori medi della deviazione standard di precisione e recall rispetto al valor medio per i diversi metodi di ranking applicati ai 10 dataset da 50 soggetti.

di interazione significative senza ricorrere alla costruzione delle reti di interazioni geniche. Si confrontino, infatti, le curve ROC relative al metodo basato su confronto della correlazione (in rosa) e al test SAM (in nero) riportate nel grafico di sinistra in figura 5.10: la prima è caratterizzata da valore di AUC medio praticamente identico a quello della seconda (0.66 contro 0.67) e presenta una veloce diminuzione dei valori di precisione in corrispondenza di valori di recall superiori rispetto alla curva ROC del test SAM.

Una riduzione significativa delle prestazioni si osserva per i metodi basati sulla misura di correlazione parziale e in modo particolare per quelli basati sul confronto della topologia delle reti di interazioni geniche. La misura di correlazione parziale, infatti, è molto sensibile alla riduzione del numero di campioni disponibili. Nei casi in cui il numero di variabili geniche G è di gran lunga superiore al numero di campioni N ($N \ll G$) per il calcolo della matrice di correlazione parziale si ricorre all'utilizzo dello stimatore presentato nella sezione 2.4.1: le prestazioni di tale stimatore peggiorano al diminuire del numero di campioni rendendo di conseguenza poco affidabili i valori di correlazione parziale calcolati.

La scarsa robustezza della misura di correlazione parziale ad una riduzione del numero di campioni è messa in luce anche dalla variabilità delle prestazioni dei metodi basati sul confronto della topologia delle reti di interazioni geniche. In figura 5.10 sono rappresentati con linea tratteggiata i valori minimi e massimi di precisione ottenuti, in corrispondenza dei diversi valori di recall, con i diversi metodi di ranking nei 10 dataset e in tabella 5.4 sono riportati i valori medi della deviazione standard di

precisione e recall rispetto al valor medio per i diversi metodi di ranking. La massima variabilità di precisione e recall si osserva proprio in corrispondenza dei metodi basati sul confronto delle reti costruite mediante correlazione parziale. Questa variabilità è molto probabilmente dovuta al fatto che la costruzione delle reti nei 10 diversi dataset dà luogo a reti di interazioni geniche con struttura variabile sia per la classe A che per la classe B: mediamente, infatti, si osserva una variazione circa del 4% degli archi individuati a partire dai dati di espressione dei diversi dataset. Questa variabilità della struttura delle reti è anch'essa imputabile alla sensibilità della misura di correlazione parziale ad una riduzione dei campioni. La struttura delle reti costruite nei vari dataset è invece molto più conservata nel caso della correlazione semplice (si osserva una variazione media solo dello 0.4% degli archi) e questo si traduce in una minor variabilità delle prestazioni dei metodi basati su confronto delle reti.

5.3.2 Stabilità delle liste

Le prestazioni dei metodi di ranking possono essere valutate non soltanto relativamente alla precisione con cui consentono di individuare i geni biomarcatori, ma anche in base alla loro capacità di ordinare i geni in liste simili a partire da versioni differenti dello stesso dataset. In questo contesto vengono considerate versioni differenti dello stesso dataset le 10 partizioni da 50 soggetti ottenute a partire dal dataset iniziale e rappresentative delle stesse classi di malattia. La robustezza delle misure per il ranking a variazioni della composizione dei dataset e la conseguente capacità dei diversi metodi di individuare liste ordinate simili, dette anche liste "stabili", è un presupposto fondamentale per poter considerare affidabili le liste di geni ottenute.

Le misure comunemente utilizzate per valutare la stabilità di due liste ordinate di geni si suddividono in *distance-based* e *set-based* [5]: alla prima categoria appartiene, ad esempio, la distanza di Canberra [73], alla seconda appartengono invece le misure che prendono in considerazione l'intersezione tra le liste di geni. In questo lavoro si è adottata una misura di stabilità di tipo set-based. In particolare, siano l_i e l_j due diverse liste ordinate di geni ottenute applicando lo stesso metodo di ranking ai due diversi dataset i e j e si considerino i geni che occupano le prime k posizioni delle due liste,

indicati rispettivamente con l_i^k e l_j^k. Si definisce stabilità delle due liste, relativamente alle prime k posizioni, il rapporto tra la cardinalità dell'intersezione di l_i^k e l_j^k e quella dell'unione:

$$\textit{Stabilità}_k = \frac{|l_i^k \cap l_j^k|}{|l_i^k \cup l_j^k|},$$

dove $|\cdot|$ indica la cardinalità di un insieme. La stabilità tra due liste assume valori nell'intervallo $[0, 1]$, in particolare assume valore massimo quando le prime k posizioni delle due liste sono occupate dagli stessi geni e intersezione e unione di l_i^k e l_j^k coincidono. Per qualsiasi coppia di liste di geni ordinate, quando si prendono in considerazione le liste complete, per $k = 900$, la stabilità assume valore unitario poiché sia l'intersezione che l'unione coincidono con la lista completa di geni.

Per non limitare l'analisi ai geni che occupano le prime k posizioni, con k fissato, è possibile valutare la stabilità delle liste per tutti i possibili valori k, ovvero per $k = 1, \ldots, 900$. Di particolare interesse è la valutazione della stabilità delle liste per valori relativamente bassi di k, ovvero in corrispondenza delle posizioni iniziali delle liste in cui compaiono i geni considerati significativi: valori elevati di stabilità per k "piccoli" indicano robustezza del metodo di ranking che è in grado di individuare le stesse variabili significative anche a partire da dataset differenti.

Per l'analisi della riproducibilità dei risultati ottenuti con i diversi metodi di ranking sono state considerate, per ciascun metodo, le 10 liste ottenute a partire dai diversi dataset ed è stata calcolata la stabilità, al variare di k, per tutte le 45 diverse possibili coppie di liste. Per una valutazione complessiva delle prestazioni è stato poi considerato il valor medio della stabilità tra tutte le possibili coppie: in figura 5.11 è riportata la stabilità media calcolata per ciascun metodo di ranking al variare di k.

Poiché i geni definiti biomarcatori sono 185, per l'analisi della riproducibilità dei risultati ottenuti con i diversi metodi di ranking è interessante, in particolar modo, confrontare la stabilità media delle liste in corrispondenza delle posizioni dalla prima alla duecentesima circa, ovvero in corrispondenza delle posizioni che, idealmente, dovrebbero essere occupate per la maggior parte da geni biomarcatori e che, in ogni caso, sono occupate dai geni considerati significativi secondo le diverse misure adottate.

Il metodo di riferimento, basato su test SAM, è caratterizzato dalla presenza di due

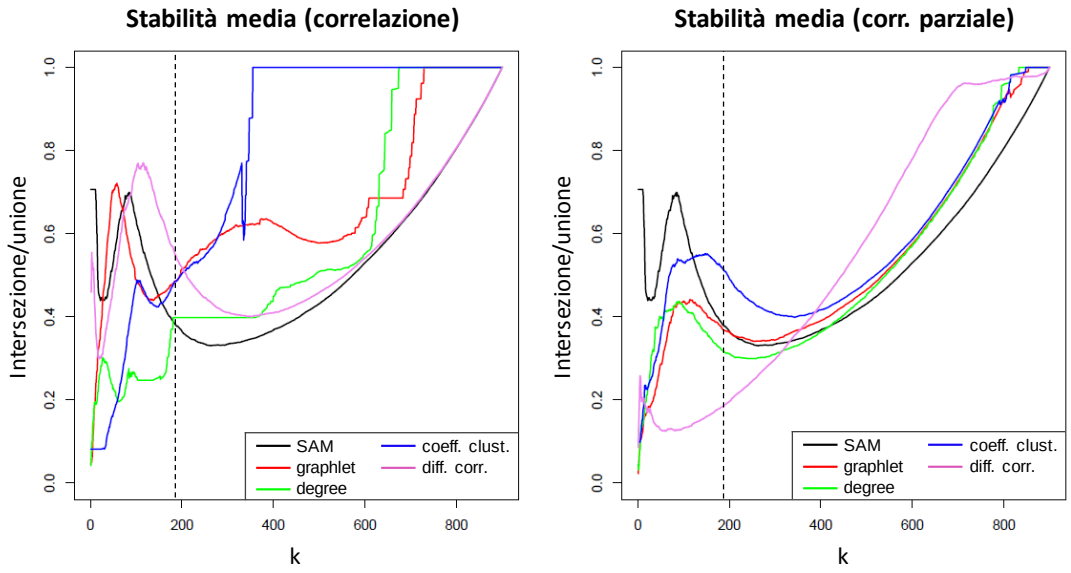


Figura 5.11 Stabilità media, al variare di k, delle coppie di liste ottenute a partire dai 10 dataset da 50 campioni utilizzando i diversi metodi di ranking proposti. Le porzioni di curva a sinistra della linea tratteggiata in corrispondenza di $k = 185$ sono quelle più interessanti per la valutazione della stabilità.

picchi nella curva di stabilità media nel range dei valori di k di interesse: il primo picco indica una stabilità pari a 0.71 per le prime $k = 11$ posizioni delle liste, il secondo, invece, indica una stabilità pari a 0.69 in corrispondenza delle prime 85 posizioni delle liste ottenute nei 10 diversi dataset.

Dall'osservazione del grafico di destra in figura 5.11 si può dedurre che i metodi che utilizzano come misura di similarità la correlazione parziale non offrono buone prestazioni in termini di stabilità delle liste e riproducibilità dei risultati. Per i k nel range di interesse la stabilità media di nessuno dei metodi supera il valore di 0.55: si osserva al massimo una corrispondenza media del 55% dei geni che occupano le prime 200 posizioni delle liste ordinate ottenute. Il fatto di ottenere risultati molto diversi in termini di ordinamento delle variabili nei 10 dataset riconferma le scarse prestazioni della misura di correlazione parziale in presenza di un numero limitato di campioni.

Risultati migliori in termini di stabilità delle liste si ottengono con i metodi di ranking basati sulla misura di correlazione semplice che, come già osservato nella sezione precedente, è caratterizzata da una maggior robustezza alla riduzione del

numero di campioni disponibili. In particolare sia il metodo basato su graphlet che quello basato sul confronto diretto dei valori di correlazione tra geni presentano dei picchi di stabilità che superano quello relativo al test SAM.

Nonostante la robustezza della misura di correlazione semplice che consente di ottenere reti di interazioni geniche caratterizzate da ridotta variabilità nei diversi dataset (0.4% degli archi), non tutti i metodi basati sul confronto della topologia delle reti ricostruite mediante correlazione presentano dei picchi di stabilità. I metodi basati sul grado di connettività e sul coefficiente di clustering, infatti, esibiscono scarse prestazioni in termini di riproducibilità dei risultati sui 10 dataset: entrambi i metodi sono caratterizzati da stabilità media inferiore a 0.5 in tutto il range di valori di k di interesse. Il fatto che la stabilità media del metodo basato su coefficiente di clustering raggiunga il valore unitario per valori di k molto inferiori a 900 non è indicativo della robustezza della misura. Questo particolare andamento è infatti spiegato dal fatto che tale metodo definisce non più di 270 diversi score per i 900 geni e dunque, per ciascuno dei 10 dataset, classifica i geni in liste di al più 270 posizioni. Il raggiungimento di stabilità unitaria per valori di $k \approx 400$ è dovuto al fatto che, nel confronto delle diverse liste, per $k \geq 400$ sono già stati presi in considerazione tutti i geni ed è come confrontare tra loro le liste complete. Considerazioni analoghe valgono per il metodo basato sul grado di connettività.

La variabilità delle liste ordinate ottenute con i metodi basati su coefficiente di clustering e grado di connettività, anche in corrispondenza di reti di interazioni geniche con struttura molto conservata nei 10 dataset, mette in evidenza l'inadeguatezza di tali misure nel quantificare in modo attendibile le variazioni di interazione tra geni. Si tratta infatti di misure che mettono a confronto proprietà topologiche molto semplici, come il grado di un nodo o la connettività dei suoi vicini, le quali sono particolarmente suscettibili anche a minime variazioni della struttura delle reti costruite nei diversi dataset.

Il picco di stabilità pari a 0.72 in corrispondenza di $k = 58$ evidenzia, al contrario, la robustezza del metodo basato su graphlet. Il fatto che ben il 72% delle prime 58 variazioni di interazione maggiormente significative coinvolga gli stessi geni nei 10

	picco di stabilità (correlazione)			
	k	*Stabilità*$_k$	*Precisione*$_k$	*Recall*$_k$
test SAM	85	0.70	0.92	0.42
graphlet	58	0.72	0.86	0.27
correlazione	115	0.77	0.82	0.51

Tabella 5.5 Valori di stabilità, precisione e recall medie in corrispondenza dei valori di k per i quali si osserva il picco di stabilità per i metodi basati su test SAM, graphlet applicati alle reti costruite mediante correlazione e confronto diretto dei valori di correlazione applicati ai dataset da 50 campioni.

dataset è riconducibile all'affidabilità di una misura che si basa sul confronto della partecipazione dei nodi a strutture complesse quali sono i graphlet e le loro orbite.

Prestazioni ancora migliori in termini di stabilità delle liste si ottengono ricorrendo al metodo basato su confronto diretto dei valori di correlazione tra coppie di geni che presenta per $k = 115$ un picco di stabilità pari a 0.77. Sebbene indicativa della robustezza delle misure di ranking utilizzate, la possibilità di ottenere liste ordinate stabili, almeno in corrispondenza delle prime posizioni, è poco significativa se tra i geni con punteggio più elevato non compaiono i geni biomarcatori. Può dunque essere utile confrontare per i diversi metodi i valori medi di precisione e recall in corrispondenza dei valori di k per i quali si osserva il picco di stabilità. Tali valori sono riportati in tabella 5.5 per i metodi di ranking basati su test SAM, graphlet applicati alle reti costruite mediante correlazione e confronto dei valori di correlazione semplice. Come si può osservare tutti e tre i metodi sono caratterizzati da una buona precisione media in corrispondenza del picco di stabilità. Il basso valore di recall del metodo basato su graphlet, pari a 0.27 e inferiore rispetto agli altri due metodi, è dovuto al fatto che il picco di stabilità si osserva per un valore di k relativamente basso.

5.3.3 25 campioni

Visti i buoni risultati ottenuti applicando i metodi basati su graphlet e su confronto diretto dei valori di correlazione ai dataset da 50 soggetti, si è pensato di testare i metodi di ranking riducendo ulteriormente il numero di campioni da 50 a 25. Sono stati

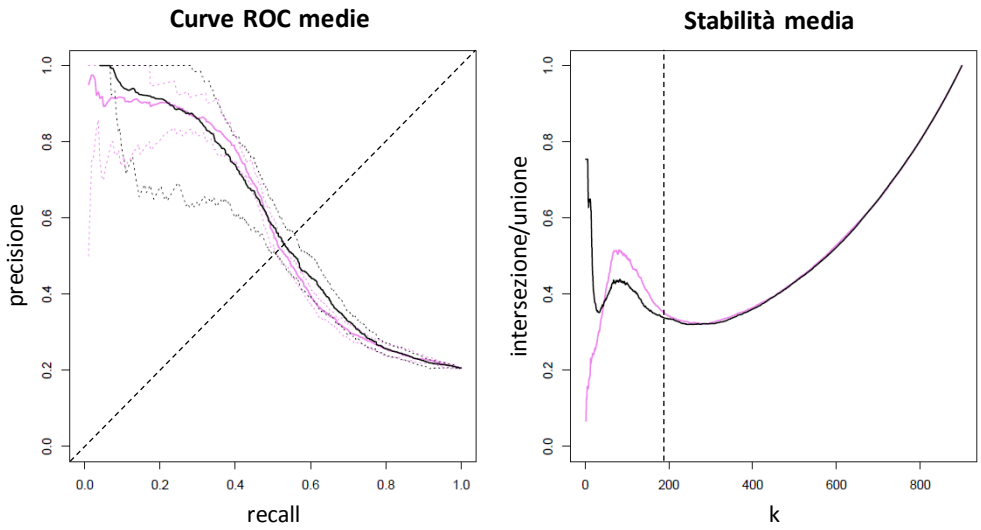


Figura 5.12 Curve ROC medie e stabilità media delle liste ottenute applicando il test SAM e il metodo basato su confronto diretto dei valori di correlazione semplice ai 10 dataset da 25 campioni.

dunque creati 10 dataset costituiti da due classi A e B di 25 soggetti ciascuna ed è stata riproposta l'applicazione dei diversi metodi di ranking, preceduta dalla costruzione delle reti di interazioni geniche.

In presenza di un numero così limitato di campioni, anche tecniche di costruzione delle reti semplici come quelle basate su confronto pair-wise dei profili di espressione genica mostrano i loro limiti. Le reti costruite, sia mediante correlazione che mediante correlazione parziale, per le due classi A e B partendo da profili di espressione costituiti da soli 25 campioni sono molto sparse, presentano cioè un numero estremamente ridotto di archi. Per queste loro caratteristiche le reti A e B costruite per i diversi dataset sono difficilmente confrontabili in termini di proprietà topologiche locali, rendendo dunque inapplicabili i metodi basati su confronto delle reti.

Il metodo basato su confronto diretto dei valori di correlazione tra coppie di geni esibisce invece buonissime prestazioni anche in corrispondenza dell'ulteriore riduzione di campioni sia in termini di precisione e recall che in termini di stabilità delle liste. In figura 5.12 sono messe a confronto le curve ROC medie e la stabilità media delle liste ottenute con metodo basato su test SAM e su massima variazione di correlazione semplice e in tabella 5.6 sono riportati i valori di stabilità, precisione e recall medie in

	picco di stabilità			
	k	*Stabilità*$_k$	*Precisione*$_k$	*Recall*$_k$
test SAM	83	0.44	0.79	0.36
correlazione	81	0.52	0.83	0.36

Tabella 5.6 Valori di stabilità, precisione e recall medie in corrispondenza dei valori di k per i quali si osserva il picco di stabilità per i metodi basati su test SAM e confronto diretto dei valori di correlazione applicati ai dataset da 25 campioni.

corrispondenza del picco di stabilità osservato per i due metodi. Come si può osservare i metodi esibiscono prestazioni simili in termini di precisione e recall (AUC media pari a 0.59 per il test SAM e 0.58 per il metodo basato su correlazione), mentre il metodo basato su correlazione è caratterizzato da un picco di stabilità superiore in corrispondenza quasi dello stesso valore di k per cui si osserva il picco del test SAM.

5.4 Riepilogo dei risultati

L'analisi dei risultati proposta in questo capitolo consente di trarre alcune conclusioni circa le prestazioni dei nuovi metodi di ranking proposti. In particolare, confrontando le prestazioni con quelle del metodo di riferimento basato su test SAM, si evince che:

- Sebbene le prestazioni del test SAM siano complessivamente migliori di quelle dei nuovi metodi di ranking proposti, alcuni di essi (come ad esempio i metodi basati su graphlet e su confronto diretto dei valori di correlazione semplice tra geni) sono comunque caratterizzati da buone prestazioni e, in alcuni casi, danno risultati migliori del test SAM (ad esempio in termini di stabilità delle liste). Questa evidenza sottolinea che il criterio di ranking basato sulle variazioni di interazione tra geni può dare buoni risultati se si dispone di una misura adeguata e sufficientemente robusta per la quantificazione di tali variazioni.

- I metodi che utilizzano come misura di similarità tra geni la correlazione semplice sono caratterizzati da una maggior robustezza al rumore e alla riduzione di campioni rispetto a quelli che utilizzano la correlazione parziale:

- la drastica riduzione di prestazioni del metodo basato sulla massima variazione di correlazione tra coppie di geni utilizzando la misura di correlazione parziale piuttosto che la correlazione semplice è imputabile all'elevata sensibilità di quest'ultima al rumore;
- la variabilità delle reti di interazioni geniche costruite mediante correlazione parziale per i 10 dataset da 50 soggetti e la diminuzione delle prestazioni dei metodi basati su confronto della topologia di tali reti è dovuta alla scarsa robustezza della misura alla riduzione del numero di campioni disponibili.

• I metodi basati su confronto delle reti di interazioni geniche esibiscono prestazioni differenti a seconda della particolare misura adottata per la descrizione delle proprietà topologiche locali delle reti:

- nessuno dei metodi è applicabile ai dataset da 25 soggetti a causa dei limiti nella costruzione delle reti di interazioni geniche (sia mediante correlazione che mediante correlazione parziale) in presenza di un numero così limitato di campioni;
- per quanto riguarda i dataset da 50 soggetti a causa della maggior sensibilità della misura di correlazione parziale, rispetto alla correlazione semplice, alla riduzione del numero di campioni e della conseguente variabilità delle reti costruite per i vari dataset nessuna delle tre diverse misure di confronto della topologia locale dà buoni risultati utilizzando come misura di similarità la correlazione parziale;
- il metodo basato su confronto del *coefficiente di clustering* dei nodi non dà buoni risultati nemmeno quando applicato ai dataset più numerosi evidenziando l'inadeguatezza della misura nel rilevare le variazioni di interazione significative che coinvolgono i geni biomarcatori;
- il metodo basato sul confronto del *grado di connettività* dei nodi è caratterizzato da buone prestazioni quando applicato ai dataset più numerosi (soprattutto se utilizzato per il confronto delle reti costruite mediante correlazione parziale), tuttavia l'elevata variabilità delle liste ottenute mediante

tale metodo per i dataset da 50 soggetti evidenzia la scarsa robustezza della misura;

- il metodo basato su *graphlet* è quello che, complessivamente, dà risultati migliori tra i metodi basati su confronto della topologia poiché esibisce buone prestazioni sia per i dataset da 500 soggetti sia per quelli da 50, sia in termini di precisione e recall che in termini di stabilità delle liste.

- I metodi basati sulla massima variazione di correlazione tra coppie di geni sono caratterizzati da prestazioni estremamente differenti a seconda che si utilizzi come misura di similarità la correlazione semplice o la correlazione parziale:

  - confrontando i valori di correlazione parziale tra coppie di geni si ottengono pessimi risultati a causa dell'elevata sensibilità della misura al rumore (le tecniche basate su confronto della topologia delle reti consentono di ottenere prestazioni migliori poiché l'introduzione della soglia di significatività limita gli effetti del rumore);
  - il metodo basato su massima variazione di correlazione semplice tra coppie di geni è quello che, in generale, esibisce le migliori prestazioni tra i nuovi metodi di ranking proposti. Tale metodo è caratterizzato da prestazioni migliori di quelli basati su confronto delle reti evidenziando l'importanza di tenere in considerazione il peso e il segno della correlazione che misura le relazioni tra coppie di geni. Inoltre tale metodo consente di ottenere buoni risultati anche in seguito ad una riduzione del numero di campioni disponibili ed è l'unico metodo caratterizzato da buone prestazioni (confrontabili con quelle del test SAM in termini di precisione e recall e superiori in termini di stabilità delle liste) anche in presenza di soli 25 campioni per ciascun profilo di espressione genica.

6

Conclusioni e sviluppi futuri

L'identificazione di geni biomarcatori a partire dai dati di espressione ottenuti mediante microarray è uno dei problemi ancora aperti e più dibattuti nell'ambito della bioinformatica. Nonostante l'elevato contenuto informativo, infatti, l'interpretazione e l'elaborazione di dati di questo tipo è resa complessa da molteplici fattori come ad esempio la rumorosità dei dati e la disponibilità di un limitato numero di campioni. Negli ultimi anni sono state proposte numerose tecniche per la selezione dei biomarcatori, che però tengono conto esclusivamente di variazioni nel livello di espressione genica.

In questo lavoro, invece, sono stati introdotti dei nuovi metodi che integrano informazioni sulle interazioni geniche, con l'intento di migliorare le prestazioni delle tecniche finora utilizzate. Le alterazioni dell'attività genica e dei meccanismi di controllo dell'espressione indotte dalle malattie genetiche si ripercuotono sulle relazioni di interazione tra geni che stanno alla base della rete di regolazione. La possibilità di tenere in considerazione tali relazioni può dunque essere rilevante ai fini dell'identificazione dei geni biomarcatori che caratterizzano la patologia analizzata. In quest'ottica, il problema principale che è stato affrontato è quello di individuare delle misure robuste per valutare il grado di correlazione tra le variabili geniche, ma soprattutto per quantificare le variazioni di interazione tra geni al fine di identificare quelle più significative a carico dei geni biomarcatori.

I nuovi metodi di ranking proposti sono stati applicati ai dati di espressione simulati di due classi di popolazione di 500 soggetti ciascuna. È dunque stato possibile valutare, oltre alla capacità delle diverse misure di individuare i geni biomarcatori, anche la loro robustezza e la riproducibilità dei risultati ottenuti in corrispondenza ad una riduzione del numero di campioni disponibili per ciascun profilo simulato.

La prima fase del lavoro è stata dedicata all'analisi della correlazione tra le variabili geniche delle due classi di popolazione simulate, e in particolare alla costruzione delle reti di interazioni geniche per le due classi mediante confronto pair-wise dei profili

di espressione. L'analisi delle proprietà topologiche delle reti ottenute, svolta per confrontare le reti costruite per le due classi di soggetti, ha messo in evidenza alcune delle limitazioni della tecnica adottata per l'identificazione delle interazioni geniche.

Le proprietà topologiche globali delle reti ottenute sia mediante correlazione che mediante correlazione parziale non riproducono esattamente quelle della rete di regolazione simulata, riconfermando il fatto che tecniche basate sul confronto pair-wise dei profili di espressione non possano essere considerate alla stregua di vere e proprie tecniche di reverse engineering per inferire sulla rete di regolazione originale. Tuttavia la ricostruzione della rete originale esula dagli scopi di questo lavoro.

Ai fini dell'applicabilità dei metodi di ranking è invece fondamentale la similarità della struttura globale riscontrata nelle reti ottenute per le due classi di soggetti. Tale caratteristica garantisce che le differenze nella topologia locale individuate mediante le diverse misure di ranking non siano dovute a differenze nella struttura globale delle reti quanto piuttosto ad una effettiva variazione delle interazioni geniche.

Alcuni limiti delle tecniche di identificazione delle interazioni geniche sono messi in luce dall'insolita presenza di nodi associati a geni biomarcatori e isolati nelle reti di entrambe le classi di soggetti: per come sono stati definiti, i biomarcatori dovrebbero presentare un elevato numero di connessioni almeno in una delle due classi. Questa difficoltà nell'identificazione delle interazioni geniche significative si ripercuote sulle prestazioni dei metodi di ranking basati su confronto della topologia locale ed evidenzia la sensibilità dell'analisi delle interazioni geniche al rumore di cui sono affetti i dati. Già in presenza di 50 campioni le reti costruite mediante correlazione parziale hanno struttura variabile evidenziando la sensibilità della misura a variazioni nel dataset. Riducendo ulteriormente il numero di campioni a 25 anche la costruzione mediante correlazione semplice mostra i suoi limiti: le reti costruite sono molto sparse e difficilmente confrontabili in termini di proprietà topologiche locali, rendendo inapplicabili i metodi di ranking basati su confronto delle reti.

Nella seconda fase di lavoro le diverse misure di ranking proposte sono state applicate ai dati simulati: i geni sono stati ordinati sulla base delle variazioni delle proprietà topologiche locali nell'intorno dei nodi ad essi associati nelle reti o sulla base della

massima variazione di correlazione o correlazione parziale con gli altri geni. Confrontando le prestazioni dei diversi metodi tra loro e con quelle del metodo classico per la selezione dei geni differenzialmente espressi basato su test SAM sono emerse alcune considerazioni interessanti circa le misure di correlazione adottate e circa le diverse misure di ranking proposte.

In primo luogo la misura di correlazione parziale, sebbene consenta di individuare con maggior affidabilità relazioni di interazione diretta tra le variabili geniche, si è rivelata, in questo contesto, non adeguata per una quantificazione robusta delle variazioni delle interazioni geniche significative. Oltre ad esibire un'elevata sensibilità al rumore e una scarsa robustezza ad una riduzione del numero di campioni disponibili, la misura di correlazione parziale è in teoria in grado di quantificare le variazioni di interazione diretta tra geni. La correlazione semplice, invece, è in grado di dare una caratterizzazione più completa delle variazioni delle modalità di interazione tra due geni, rilevando anche il contributo legato ad una variazione delle interazioni delle due variabili con i restanti geni.

Tra le nuove misure proposte per il ranking dei geni quelle basate sul confronto del grado di connettività o del coefficiente di clustering dei nodi non esibiscono buone prestazioni e non risultano essere misure adeguate per l'identificazione dei biomarcatori. I risultati ottenuti applicando il metodo basato su coefficiente di clustering al dataset da 500 soggetti, non soddisfacenti in termini di precisione e recall, mettono in luce l'inadeguatezza della misura: la quantificazione delle variazioni di interazione tra i geni direttamente connessi ad uno stesso gene nelle due reti non consente di identificare le variazioni di interazione significative che coinvolgono i geni biomarcatori.

Il metodo basato sul grado di connettività, invece, consente di ottenere dei buoni risultati se applicato al dataset da 500 soggetti, tuttavia l'elevata variabilità delle liste di biomarcatori ottenute per i dataset da 50 soggetti evidenzia la scarsa robustezza della misura e la sua elevata sensibilità a variazioni nei dataset.

I metodi basati su confronto della partecipazione dei nodi ai graphlet e alle loro orbite e sulla massima variazione di correlazione consentono invece di ottenere risultati decisamente migliori. Entrambi i metodi, infatti, superano il test SAM in termini di

riproducibilità dei risultati e stabilità delle liste ottenute. Il metodo basato su variazione di correlazione, inoltre, esibisce prestazioni confrontabili con quelle del test SAM in termini di precisione e recall e, poiché non richiede la costruzione delle reti di interazioni geniche ma confronta indistintamente tutti i valori di correlazione calcolati nelle due classi di soggetti, risulta applicabile anche ai dataset da 25 campioni dove restituisce ottimi risultati.

Queste evidenze sottolineano che un criterio di ranking basato sulle variazioni di interazione tra geni può dare ottimi risultati se si dispone di una misura adeguata e sufficientemente robusta per la quantificazione di tali variazioni. In particolare, la possibilità di tenere in considerazione la partecipazione dei nodi ai moduli di interazione rappresentati dai graphlet consente di ottenere una buona caratterizzazione delle modalità di interazione genica e di conseguenza una quantificazione robusta delle loro variazioni che consente di individuare con una buona precisione quelle che coinvolgono i geni biomarcatori. Anche le informazioni sul segno e sul peso associato ai valori di correlazione e sulle differenze di correlazione tra geni nelle due classi di soggetti a confronto sono di fondamentale importanza per l'identificazione robusta dei geni biomarcatori.

Complessivamente i risultati ottenuti sono molto promettenti e, oltre ad evidenziare la validità di un criterio di ranking che individua come biomarcatori i geni che hanno subito la massima variazione di interazione con gli altri geni, pongono le basi per possibili sviluppi futuri e miglioramenti dei metodi e delle diverse misure proposte.

Per testare ulteriormente la capacità dei nuovi metodi di ranking proposti nell'identificare i geni biomarcatori potrebbe essere interessante applicarli agli stessi dati simulati utilizzati in questo lavoro, ricorrendo però a ulteriori misure di similarità tra profili di espressione genica. Si è infatti osservato che le prestazioni delle diverse misure di ranking sono fortemente condizionate dalla misura di similarità utilizzata per valutare il grado di correlazione tra variabili e per costruire le reti di interazioni geniche. Un tentativo è già stato fatto utilizzando come misura di similarità la mutua informazione tra profili di espressione genica. Le prestazioni dei diversi metodi di ranking sono confrontabili con quelle ottenute mediante correlazione semplice, tuttavia la misura

di similarità basata su mutua informazione richiede un numero elevato di campioni per dare una quantificazione robusta del grado di interazione tra variabili geniche e le prestazioni dei diversi metodi di ranking peggiorano velocemente, come osservato per la correlazione parziale, in corrispondenza ad una riduzione del numero di campioni disponibili.

Un'altra possibilità è quella di ricorrere a metodi di analisi delle interazioni geniche diversi da quelli basati su confronto pair-wise dei profili di espressione, i quali tengono in considerazione soltanto le relazioni tra coppie di variabili trascurando invece pattern di interazione più complessi. Utilizzando metodi basati su modello è possibile ottenere una caratterizzazione più completa delle relazioni di interazione tra le variabili, che consente di descrivere, ad esempio, come i regolatori di uno stesso gene interagiscono tra loro nell'esplicare la loro azione regolatoria. Da ciò potrebbero trarre vantaggio i metodi basati su graphlet che tengono in considerazione la partecipazione dei nodi a diverse strutture modulari anche molto articolate. Tuttavia è importante sottolineare che i metodi basati su modello sono di difficile applicazione in presenza di un numero elevato di variabili e di un numero limitato di campioni a disposizione.

Dati i buoni risultati ottenuti ricorrendo alle misure di variazione delle interazioni geniche basate su graphlet e sulla massima differenza di correlazione, un'interessante evoluzione dei metodi proposti in questo lavoro potrebbe consistere nell'integrazione di queste due misure. Per combinare le diverse informazioni sulle variazioni di interazione prese in considerazione da tali misure, una possibilità è quella di costruire delle reti di interazioni geniche pesate, selezionando le relazioni significative e associando un peso agli archi individuati corrispondente al valore di correlazione calcolato per la coppia di geni. Confrontando la partecipazione dei nodi ai graphlet e alle loro orbite, tenendo contemporaneamente in considerazione i pesi associati agli archi delle reti è possibile integrare informazioni sulla variazione del valore di correlazione tra coppie di geni nelle due classi e sulla variazione di partecipazione alle strutture modulari rappresentate dai graphlet.

I risultati ottenuti e le considerazioni emerse dal confronto dei metodi di ranking proposti con quelle del metodo classico basato su test SAM suggeriscono infine la

possibilità di pensare a nuovi metodi per l'identificazione dei geni biomarcatori che integrino informazioni sulla variazione del livello di espressione e sulla variazione delle interazioni geniche. Il metodo per la selezione dei geni differenzialmente espressi basato su test SAM esibisce infatti prestazioni migliori in termini di precisione e recall, mentre i metodi basati su graphlet e sulla massima variazione di correlazione superano il test SAM in termini di stabilità delle liste ottenute. Il raggiungimento di buone prestazioni in entrambi i frangenti potrebbe essere ottenuto proprio mediante l'integrazione delle informazioni di diversa natura prese in considerazione dai diversi metodi. In questo modo è infatti possibile dare una più completa caratterizzazione del manifestarsi della patologia analizzata e delle alterazioni dell'attività genica da essa indotte che si ripercuotono sia sul livello di espressione delle singole variabili geniche sia sulle relazioni di interazione che intercorrono tra i geni.

A

Correzione per test multipli

Come accennato nella sezione 2.4.3, il controllo della FP-rate del singolo test statistico non è un criterio adeguato ad una situazione di test statistici multipli. Si supponga di essere disposti a commettere errori di tipo falso positivo con probabilità α e si scelga come soglia di confidenza θ per i singoli test quella corrispondente al livello di significatività α. Sia E il numero di potenziali archi della rete per i quali è necessario svolgere un test statistico e si supponga di conoscere il numero vero di archi E_0 che non sono presenti nella rete, corrispondente al numero di coppie di geni realmente non interagenti. Considerando i test statistici come E prove ripetute e indipendenti (prove di Bernoulli), il numero atteso di errori di tipo falso positivo commessi complessivamente su tutti i test svolti è dato da $E[\#FP] = E_0\alpha$. La probabilità di commettere errori di tipo falso positivo sugli E test è dunque notevolmente aumentata rispetto al valore desiderato α: è necessario introdurre un criterio per la determinazione del livello di significatività che consenta di controllare il numero di errori di tipo falso positivo commessi globalmente sugli E test.

In questo lavoro si è scelto di adottare un criterio di correzione per test multipli basato su false discovery rate (FDR) [52], definita come:

$$FDR = E\left[\frac{\#FP}{S_\alpha}\right] = \frac{E_0\alpha}{S_\alpha},$$

dove S_α indica il numero di archi in corrispondenza dei quali l'ipotesi nulla viene rifiutata e che vengono selezionati nella rete di interazioni geniche. Un controllo del valore della false discovery rate consente dunque di scegliere il livello di significatività α in modo tale che la percentuale di errori di tipo falso positivo sul totale delle relazioni geniche selezionate non superi una certa soglia Q. Il procedimento da seguire per determinare il livello di significatività che consente di controllare la false discovery rate al valore desiderato Q può essere così schematizzato:

1. Si costruisce la lista dei p-value ordinati in ordine crescente $p_1, p_2, \ldots, p_E$ e la lista dei corrispondenti archi $e_1, e_2, \ldots, e_E$.

2. Si calcola il valore di FDR per ciascun p-value come $FDR_i = (E_0 p_i)/i$; avendo ordinato i p-value in ordine crescente, infatti, in corrispondenza ad una scelta del livello di significatività pari a p_i, il numero di relazioni selezionate, dato dal numero di archi con p-value minore o uguale a p_i, è proprio i.

3. Si definisce l'indice i_Q come il più grande i per il quale $FDR_i \leq Q$.

4. Si rifiuta l'ipotesi nulla per tutti gli archi $e_1, e_2, \ldots, e_{i_Q}$ il cui p-value è minore o uguale a p_{i_Q}.

Il livello di significatività che consente di controllare la FDR al valore desiderato Q sarà dunque dato da $\alpha = p_{i_Q}$ e la soglia di confidenza θ sarà il valore di correlazione (o correlazione parziale), in modulo, associato alla coppia di geni con p-value p_{i_Q}. Per il calcolo della false discovery rate è necessario conoscere il numero di coppie di geni realmente non interagenti, E_0. Poiché, ovviamente, non si dispone di questa informazione è possibile approssimare E_0 con il numero complessivo di archi E: si tratta di una scelta molto conservativa che tende a sovrastimare E_0 e, di conseguenza, il numero atteso di errori di tipo falso positivo. In alternativa è possibile stimare E_0 a partire dai dati.

Sia E il numero totale di coppie di geni, si indica con E_1 il numero di coppie di geni realmente correlati e con E_0 le restanti coppie di geni non correlati. La stima di E_0 può essere ottenuta a partire dai dati riproponendo la procedura di selezione delle relazioni significative con una particolare scelta del livello di significatività α. Ciascuno degli E potenziali archi della rete viene testato, sottoponendo a test statistico il valore di correlazione (o correlazione parziale) ad esso associato, scegliendo un livello di significatività λ sufficientemente elevato. In questo modo aumentano sicuramente gli errori di tipo falso positivo commessi, ma aumenta anche la probabilità di selezionare tutte le E_1 coppie di geni realmente interagenti. La soglia di confidenza θ_λ associata ad un livello di significatività λ elevato avrà infatti un valore prossimo a zero: è dunque altamente probabile che tutti i valori di correlazione delle E_1 coppie di geni realmente

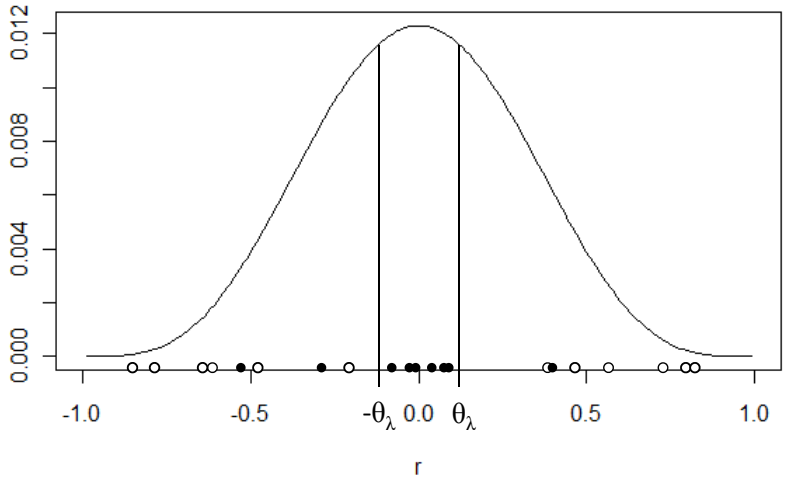


Figura A.1 In corrispondenza ad un elevato livello di significatività λ tutte le coppie di geni realmente correlate (in bianco) vengono selezionate, mentre i valori di correlazione compresi nell'intervallo $[-\theta_\lambda, \theta_\lambda]$ corrispondono a coppie di geni non correlati (in nero).

correlati cadano fuori dall'intervallo $[-\theta_\lambda, \theta_\lambda]$, come rappresentato in figura A.1, e che all'interno dell'intervallo cadano esclusivamente valori di correlazione di coppie di geni non correlati. Le coppie di geni selezionate S_λ, caratterizzate da correlazione superiore al valore di soglia θ_λ, possono essere pensate come la somma di falsi positivi, coppie selezionate ma non correlate nella realtà, e di veri positivi, coppie selezionate e realmente correlate. Il numero di falsi positivi è approssimabile con il valore atteso $E_0\lambda$, mentre il numero di veri positivi corrisponde ad una frazione $k \approx 1$ delle E_1 coppie di geni realmente correlate. In formule:

$$S_\lambda = \#FP + \#VP \approx E_0\lambda + kE_1 \approx E_0\lambda + E_1. \tag{A.1}$$

A partire dall'equazione (A.1) e sostituendo E_1 con $E - E_0$ è infine possibile stimare E_0 come:

$$\hat{E_0} = \frac{E - S_\lambda}{1 - \lambda}.$$

Per quanto riguarda la scelta del livello di significatività λ si calcola $\hat{E}_0$ per diversi valori di λ e si sceglie il λ che minimizza l'errore quadratico medio sulla stima.

B

Test statistico SAM

Per la validazione del nuovo metodo proposto per il ranking dei geni, i risultati ottenuti sono stati messi a confronto con quelli ottenuti applicando il metodo SAM (*Significance Analysis of Microarray*) proposto in [13]. Si tratta di un metodo frequentemente utilizzato per l'analisi di dati di espressione genica che rientra tra le tecniche di feature selection per la selezione di biomarcatori. Per l'applicazione del metodo SAM sono state utilizzate le funzioni implementate, in linguaggio di programmazione R, nel pacchetto siggenes scaricabile dall'archivio di Bioconductor [48].

In particolare il metodo SAM consiste nell'applicazione di un test statistico per la selezione dei geni differenzialmente espressi nelle due classi di soggetti considerate. Il test statistico utilizzato è una versione modificata del test di Student [74], o t-test, che è stato adattato per l'applicazione a dati di espressione genica ottenuti da esperimenti con microarray. Il test SAM offre, infatti, buone prestazioni quando applicato a dati di microarray poiché consente di tenere in considerazione la presenza di rumore e la disponibilità di un numero limitato di campioni che caratterizza questo tipo di dati.

Il test SAM assegna a ciascuno dei geni monitorati una variabile che definisce la variazione del livello di espressione del gene nelle due classi in rapporto alla deviazione standard dei campioni disponibili per il gene stesso. La statistica adottata consente dunque di tenere in considerazione, oltre alla variazione del livello medio di espressione genica, anche la presenza di variazioni gene-specifiche nei valori di espressione misurati. Siano $g_i^A = (g_{i1}^A, \ldots, g_{iN_A}^A)$ e $g_i^B = (g_{i1}^B, \ldots, g_{iN_A}^B)$ rispettivamente i campioni di espressione del gene g_i nella classe A e nella classe B. La variabile assegnata al gene g_i, che ne definisce la variazione relativa del livello di espressione, è data da:

$$t_{\text{i,SAM}} = \frac{\overline{g_i}^A - \overline{g_i}^B}{s_i + s_0}, \tag{B.1}$$

dove $\overline{g_i}^A$ e $\overline{g_i}^B$ sono i livelli di espressione media del gene g_i rispettivamente nella classe A e nella classe B. La variabile s_i tiene conto della dispersione gene-specifica dei valori di espressione osservati ed è definita dalla deviazione standard dei campioni

disponibili per il gene g_i nelle due classi:

$$s_i = \sqrt{a\left\{\sum_{j=1}^{N_A}\left[g_{ij}^A - \overline{g_i}^A\right]^2 + \sum_{j=1}^{N_B}\left[g_{ij}^B - \overline{g_i}^B\right]^2\right\}}, \tag{B.2}$$

con $a = (1/N_A + 1/N_B)/(N_A + N_B - 2)$.

Il termine s_0 che compare a denominatore nell'equazione (B.1) è una costante positiva che assume lo stesso valore per tutti i geni. Si tratta di un termine correttivo introdotto per minimizzare gli effetti della possibile sottostima di s_i, dovuta alla scarsità di campioni a disposizione. Il valore di s_0 viene calcolato minimizzando un coefficiente di variazione dei valori dei $t_{\text{i,SAM}}$ espressi in funzione di s_i. In questo modo si ottengono delle variabili $t_{\text{i,SAM}}$ la cui distribuzione è indipendente dal livello di espressione dei singoli geni ed è possibile confrontare tra loro le variabili calcolate per ciascun gene [13].

Il test statistico SAM è un test non parametrico che consente di selezionare i geni differenzialmente espressi senza fare alcuna ipotesi a priori sulla distribuzione dei dati. In particolare la distribuzione di t_{SAM} in ipotesi nulla viene stimata a partire dai dati stessi mediante permutazioni dei campioni di espressione genica. Permutando i campioni di espressione genica a disposizione vengono generate due classi di soggetti miste e bilanciate, caratterizzate ciascuna da un egual numero di soggetti provenienti dalle due classi originarie A e B. Le due classi così ottenute sono rappresentative di una stessa popolazione di soggetti, detta appunto popolazione mista. I valori assunti dai $t_{\text{i,SAM}}$ in questa particolare situazione corrispondono a realizzazioni in ipotesi nulla. Permutando in tutti i possibili diversi modi i dati e calcolando i valori di $t_{\text{i,SAM}}$ relativi alle classi miste così ottenute è possibile determinare una stima della distribuzione di t_{SAM} in ipotesi nulla.

A partire dalla distribuzione in ipotesi nulla e dai valori di $t_{\text{i,SAM}}$ calcolati, è possibile determinare il p-value associato a ciascuna variabile genica. Il p-value associato al gene g_i, $p\text{-}value_i$, rappresenta la probabilità che il gene g_i verifichi l'ipotesi nulla e non sia differenzialmente espresso nelle due classi di soggetti e assume valori nell'intervallo $[0, 1]$. Per ottenere un ranking dei geni confrontabile con quello proposto nelle sezioni precedenti, a ciascuna variabile genica viene associato lo score

$$S_{\text{i,SAM}} = 1 - p\text{-}value_i. \tag{B.3}$$

In questo modo geni con uno score elevato (prossimo a 1) corrispondono a geni con valori di espressione significativamente diversi nelle due classi di soggetti. Ordinando i geni secondo un ordinamento decrescente dello score ad essi associato, dunque, si ottiene una lista i cui primi posti sono occupati dai geni differenzialmente espressi. Questa lista ordinata può essere confrontata con quelle ottenute mediante i nuovi score definiti, che ordinano i geni in base alla variazione di interazioni con altri geni. Dal confronto delle liste ottenute e conoscendo a priori la lista di biomarcatori (i metodi sono stati applicati a dati di espressione simulati) è possibile valutare le prestazioni dei nuovi metodi proposti per il ranking dei geni.

Bibliografia

[1] D.J. Lockhart *et al.* "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays", *Nature Biotechnology*, vol. 14, pp. 1675-1680, 1996.

[2] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", *Science*, vol. 270, pp. 467-470, 1995.

[3] A. Fantoni, S. Bozzaro, G. Del Sal, S. Ferrari, M. Tripodi, *Biologia cellulare e genetica*, Piccin, 2009.

[4] K.E. Desrumeaux, D. Baty, P. Chames, "State of the Art in Tumor Antigen and Biomarker Discovery", *Cancers*, no. 3, pp. 2554-2596, 2011.

[5] A.L. Boulesteix, M. Slawski, "Stability and Aggregation of Ranked Gene Lists", *Briefings in Bioinformatics*, vol. 10, pp. 556-568, 2009.

[6] S.Y. Kim, "Effects of Sample Size on Robustness and Prediction Accuracy of a Prognostic Gene Signature", *BMC Bioinformatics*, vol. 10, 2009.

[7] L. Klebanov, A. Yakovlev, "How High is the Level of Technical Noise in Microarray Data?", *Biology Direct*, no. 2, 2007.

[8] J.P.A. Ioannidis *et al.*, "Repeatability of Published Microarray Gene Expression Analyses", *Nature Genetics*, vol. 41, no. 2, pp. 149-155, 2009.

[9] R.A. Irizarry *et al.*, "Multiple-Laboratory Comparison of Microarray Platforms", *Nature Methods*,vol. 2, pp. 345-350, 2005.

[10] X. Solé *et al.*, "Biological Convergence of Cancer Signatures", *PLoS ONE*, vol. 4 no. 2, 2009.

[11] X. Cui, G.A. Churchill, "Statistical Tests for Differential Gene Expression in cDNA Microarray Experiments", *Genome Biology*, vol. 4, no. 210, 2003.

[12] P. Jafari, F. Azuaje, "An Assessment of Recently Published Gene Expression Data Analyses: Reporting Experimental Design and Statistical Factors", *BMC Medical Informatics and Decision Making*, vol. 6, no. 27, 2006.

[13] V.G. Tusher, R. Tibshirani, G. Chu, "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response", *Proceedings of the National Academy of Science*, vol. 98, pp. 5116-5121, 2001.

[14] T.H. Bø, I. Jonassen, "New Feature Subset Selection Procedures for Classification of Expression Profiles", *Genome Biology*, vol. 3, no. 4, 2002.

[15] E.Yeoh, *et al.*, "Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling", *Cancer Cell*, vol. 1, pp. 133-143, 2002.

[16] A. Buness, M. Ruschhaupt, R. Kuner, A. Tresch, "Classification Across Gene Expression Microarray Studies", *BMC Bioinformatics*, vol. 10, no. 453, 2009.

[17] R. Tibshirani, L. Wasserman, "Correlation-Sharing for Detection of Differential Gene Expression", *Technical Report*, 2006.

[18] J.D. Storey, J.Y. Dai, J.T.Leek, "The Optimal Discovery Procedure for Large-Scale Significance Testing, with Apllications to Comparative Microarray Experiments", *Biostatistics*, vol. 8, no. 2, pp. 414-432, 2007

[19] V. Zuber, K. Strimmer, "Gene Ranking and Biomarker Discovery Under Correlation", *Bioinformatics*, vol. 25, no. 20, pp. 2700-2707, 2009.

[20] Y. Lai, "Genome-Wide Co-Expression Based Prediction of Differential Expressions", *Bioinformatics*, vol. 24, no. 5, pp. 666-673, 2008.

[21] J. Zhang, Y. Xiang, L. Ding, K.K. Circle, T.B. Borlawsky, H.G. Ozer, R. Jin, P. Payne, K. Huang, "Using Gene Co-Expression Network Analysis to Predict Biomarkers for Chronic Lymphocytic Leukemia", *BMC Bioinformatics*, vol. 10, 2010.

[22] N. Pržulj, "Biological Network Comparison using Graphlet Degree Distribution", *Bioinformatics*, vol. 23, pp. e177-e183, 2006.

[23] B. Di Camillo, G. Toffolo, C. Cobelli, "A Gene Network Simulator to Access Reverse Engineering Algorithms", *Annals of the New York Academy of Sciences*, no. 1158, pp. 125-142, 2009.

[24] M. Bansal, V. Belcastro, A. Ambesi, A. Impiombato, D. di Bernardo, "How to Infer Gene Networks from Expression Profiles", *Molecular System Biology*, no. 3, p. 78, 2007.

[25] R. Bellazzi, S. Bicciato, S. Cavalcanti, C. Cobelli, G.M. Toffolo, *Genomica e proteomica computazionale*, Pàtron, 2007.

[26] T.S. Gardner, J.J. Faith, "Reverse Engineering Transcriptional Control Networks", *Phisics of Life Reviews*, no. 3, pp. 65-88, 2005.

[27] N. Soranzo, G. Bianconi, C. Altafini, "Comparing Association Network Algorithms for Reverse Engineering of Large Scale Gene Regulatory Networks: Synthetic vs Real Data", *Bioinformatics*, vol. 00, no. 00, pp. 1-7, 2007.

[28] S. Liang, S. Fuhrman, R. Somogyi, "REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Networks Architectures", *Pacific Symposium on Biocomputing* 1998, no. 3, pp. 18-29, 1998.

[29] I. Shmulevich, E. Dougherty, S. Kim, W. Zhang, "Probabilistic Boolean Networks: a Rule-Based Uncertainty Model for Gene Regulatroy Networks", *Bioinformatics*, vol. 18 no. 2, pp. 261-274, 2001.

[30] N. Friedman, M. Linial, I. Nachman, D. PE'ER, "Using Bayesian Networks to Analyse Expression Data", *Journal of Computational Biology*, vol. 7, pp. 601-620, 2000.

[31] N. Friedman, “Inferring Cellular Networks Using Probabilistic Graphical Models”, *Science*, vol. 303, pp. 799-805, 2004.

[32] A. de la Fuente, P. Brazhnik, P. Mendes, “Linking the Genes: Inferring Quantitative Gene Networks from Microarray Data”, *Trends Genet*, vol. 18, pp. 395-398, 2002.

[33] P. D’haeseleer, X. Wen, S. Fuhrman, R. Somogyi, “Linear Modeling of mRNA Expression Levels During CNS Development and Injury”, *Pacific Symposium on Biocomputing*, no. 4, pp. 41-52, 1999.

[34] T.S. Gardner, D. di Bernardo, D. Lorenz, J.J. Collins, “Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling”, *Science*, vol. 301, pp. 102-105, 2003.

[35] A.J. Butte, I.S. Kohane, “Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements”, *Pacific Symposium on Biocomputing*, vol. 5, pp. 415-426, 2000.

[36] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, A. Califano, “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context”, *BMC Bioinformatics*, vol. 7, 2006.

[37] P. D’haeseleer, X. Wen, S. Fuhrman, R. Somogyi, “Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data”, *Information Processing in Cells and Tissues*, pp. 203-212, Paton, R.C., and Holcombe, M. Eds., Plenum Publishing, 1998.

[38] J. Herrero, R. Díaz-Uriarte, J. Dopazo, “An Approach to Inferring Transcriptional Regulation Among Genes from Large-Scale Expression Data”, *Comparative and Functional Genomics*, vol. 4, pp. 148-154, 2003.

[39] A. de la Fuente, N. Bing, I. Hoeschele, P. Mendes, “Discovery of Meaningful Associations in Genomic Data Using Partial Correlation Coefficients”, *Bioinformatics*, vol. 20, no. 18, pp. 3565-3574, 2004.

[40] H. Kishino, P. Waddell, "Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data", *Genome Informatics* vol. 11, pp. 83-95, 2000.

[41] P.M. Magwene, J. Kim, "Estimating Genomic Coexpression Networks Using First-Order Conditional Independence", *Genome Biology*, vol. 5, R100, 2004.

[42] J. Schäfer, K. Strimmer, "An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks", *Bioinformatics*, vol. 21, no. 6, pp. 754-764, 2005.

[43] K. Pearson, "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia", *Philosophical Transactions of the Royal Society of London*, Series A, pp. 253-318, 1896.

[44] D. Edwards, *Introduction to Graphical Modelling*, Springer, 2000.

[45] A.J. Butte, I.S. Kohane, "Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks", *Proc. AMIA. Symposium*, pp. 711-715, 1999.

[46] P. D'haeseleer, S. Liang, R. Somogyi, "Genetic Network Inference: from Co-Expression Clustering to Reverse Engineering", *Bioinformatics*, vol. 16, pp. 707-726, 2000.

[47] *The R Project for Statistical Computing*, http://www.r-project.org

[48] *Bioconductor, Open Source Software for Bioinformatics*, http://www.bioconductor.org

[49] R. Penrose, "A Generalized Inverse for Matrices", *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, pp. 406-413, 1955.

[50] H. Hotelling, "New Light on the Correlation Coefficient and Its Transforms", *Journal of the Royal Statistical Society B*, vol. 15, no. 2, pp. 193-232, 1953.

[51] L. Euler, "De progressionibus transcendentibus seu quarum termini generales algebraice dari nequeunt", *Comm. Acad. Sci. Petropolitanae*, vol. 5, pp. 36-57, 1730.

[52] Y. Benjamini, Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289-300, 1995.

[53] V. Kulasingam, E.P. Diamandis, “Strategies for Discovering Novel Cancer Biomarkers through Utilization of Emerging Technologies”, *Nature Clinical Practice*, vol. 5, no. 10, pp. 588-599, 2008.

[54] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring” , *Science*, vol. 286, pp. 531-537, 1999.

[55] Y. Saeys, I. Inza, P. Larrañaga, “A Review of Feature Selection Techniques in Bioinformatics”, *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[56] J.G. Thomas, J.M. Olson, S.J. Tapscott, *et al.*, “An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles”, *Genome Reasearch*, no. 11, pp. 1227-1236, 2001.

[57] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, “Rank Products: a Simple, Yet Powerful, New Method to Detect Differentially Regulated Genes in Replicated Microarray Experiments”, *Federation of European Biochemical Societies*, no. 573, pp. 83-92, 2004.

[58] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, “Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data”, *Bioinformatics*, vol. 16, no. 10, pp. 906-914, 2000.

[59] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, “Gene Selection for Cancer Classification Using Support Vector Machine”, *Machine Learning*, vol. 46, pp. 389-422, 2002.

[60] T. Milenković, N. Pržulj, “Uncovering Biological Network Function via Graphlet Degree Signature”, *Cancer Information*, vol. 6, pp. 257-273, Epub, 2008.

[61] T. Milenković, J. Lai, N. Pržulj, “GraphCrunch: A Tool for Large Network Analyses”, *BMC Bioinformatics*, vol. 9, no. 70, 2008.

[62] D.E. Featherstone, K. Broadie, “Wrestling with Pleiotropy: Genomic and Topological Analysis of the Yeast Gene Expression Network”, *BioEssays*, no. 24, pp. 267-274, 2002.

[63] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabási, "The Large-Scale Organization of Metabolic Networks", *Nature*, no. 407, pp. 651-654, 2000.

[64] M.M. Babu, N.M. Luscombe, L. Aravind, L. Gerstein, S. Teichmann, " Structure and Evolution of Transcriptional Regulatory Networks", *Current Opinion in Structural Biology*, no. 14, pp. 283-291, 2004.

[65] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks", *Science*, vol. 297, pp. 1551-1555, 2002.

[66] D.J. Watts, S.H. Strogatz, "Collective Dynamics of 'Small-World' Networks", *Nature*, vol. 393, pp. 440-442, 1998.

[67] S.A. Kauffman, "Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets", *Journal of Theoretical Biology*, no. 22, pp. 437-467, 1969.

[68] A.L. Barabási, R. Albert, "Emergence of Scaling in Random Networks", *Science*, no. 286, pp. 509-512, 1999.

[69] N. Pržulj, D.G. Corneil, I. Jurisica, "Modeling Interactome: Scale-Free or Geometric?", *Bioinformatics*, vol. 20, pp. 3508-3515, 2004.

[70] S.S. Shen-Orr, R. Milo, S. Mangan, U. Alon, "Network Motifs in the Transcriptional Regulation Network of Escherichia Coli", *Nature Genetics*, vol. 31, pp. 64-68, 2002.

[71] *Gene Expression Omnibus*, http://www.ncbi.nlm.nih.gov/geo/

[72] A.L. Barabási, Z. Dezső, E. Ravasz, S. Yook, Z. Oltvai, "Scale-free and Hierarchical Structures in Complex Networks", *Modeling of Complex Systems: Seventh Granada Lectures*, vol. 661, pp. 1-16, 2003.

[73] G. Jurman, S. Riccadonna, R. Visintainer, C. Furlanello, "Canberra Distance on Ranked Lists", *In Proceedings of Advances in Ranking NIPS*, pp. 22-27, 2009.

[74] L. Soliani, F. Sartore, E. Siri, *Manuale di statistica per la ricerca e la professione. Statistica univariata e bivariata parametrica e non-parametrica per le discipline ambientali e biologiche*, 2005.