Università della Calabria

UNIVERSITÀ DELLA CALABRIA

DIPARTIMENTO DI
INGEGNERIA INFORMATICA,
MODELLISTICA, ELETTRONICA
E SISTEMISTICA

DIMES

Corso di laurea triennale in

Ingegneria Informatica A.A. 2016-2017

Tesi di Laurea

"Ricerca di pattern discriminanti su reti biologiche"

Relatore

Prof. Fabio Fassetti

Candidato

Ivonne Rizzuto
mat.167058

_____ _____

Indice

Introduzione

Sin dagli anni ‘70 la ricerca si è indirizzata verso l’elaborazione di varie tecniche di sequenziamento del DNA, con l’obiettivo di arricchire, quanto più possibile, la conoscenza del genoma umano, tentando di acquisire qualsiasi tipo di informazioni sulla base dell’ordine in cui i diversi nucleotidi si susseguono nella composizione dell’acido nucleico.

Il campo medico è quello che più trarrebbe benefici dai progressi scientifici effettuati in quest’ambito, poiché la comprensione dei meccanismi cellulari e delle entità biologiche coinvolte nella loro realizzazione, come geni e proteine, permetterebbe, ad esempio, di identificare e diagnosticare malattie dovute a mutazioni o disfunzioni cellulari, sviluppando, di conseguenza, dei trattamenti che siano mirati ed innovativi.

Questo elaborato ha lo scopo di descrivere, in generale, i modelli che sono stati utilizzati per rappresentare queste entità biochimiche e le tecniche che sono state applicate per analizzarli, concludendo con la presentazione di un algoritmo impiegato proprio per evidenziare se esistano delle corrispondenze tra i profili genici di vari campioni, in modo tale da poter distinguere delle strutture molecolari che si presentino, invariate, alle medesime condizioni.

I dati sui quali è stato applicato l’algoritmo di Discriminative Pattern Mining sono delle reti bilogiche rappresentate attraverso un’organizzazione a grafo, in modo tale che da queste vengano estratti i sottografi eccezionali e quindi più significativi.

È stato però necessario, prima di poter procedere con l’esecuzione dell’analisi, attuare una fase di preprocessing dei dati stessi.

Questa consta di una serie di operazioni di pulizia e rielaborazione, atte anche al recupero di informazioni, quali i nomi dei geni presenti nei campioni ed il quadro clinico dei pazienti dai quali sono stati estratti.

È sulla base di questi parametri che verranno modellate le reti da associare al materiale genico di interesse, sulla cui elaborazione si incentra la fase di sperimentazione.

Sono stati effettuati, quindi, due differenti scenari di esperimenti, in modo tale da identificare i sottografi discriminanti per la coppia di reti analizzate per ogni set di profili genici e ottenere, così, dei risultati che permettessero di confrontare come il numero delle corrispondenze trovate differisse nelle sequenze nucleotidiche estratte da individui sani, rispetto a quelle appartenenti a soggetti malati.

Inoltre, è stato possibile rielaborare, a loro volta, queste informazioni, per mappare i geni coinvolti nelle relazioni emerse dal riscontro di queste sottostrutture, risalendo in questo modo, ai loro nomi biologici.

Nel dettaglio, nel primo capitolo si definirà il concetto di rete biologica e come questo possa essere impiegato per formalizzare le interazioni che avvengo tra gli agglomerati proteici che conducono le attività biochimiche di un essere vivente.

In particolare, ci si soffermerà sulla descrizione delle caratteristiche che riguardano le reti di interazione genica e quelle relative alle interdipendenze che si instaurano tra differenti proteine, soffermandosi sulle fonti da cui è possibile reperire i dati per la loro rappresentazione.

Nel secondo capitolo seguirà una presentazione delle principali tecniche di analisi da utilizzare su questo tipo di reti, in modo tale da comparare, tra loro, le informazioni relative a campioni di organismi appartenti a specie diverse, così come le sequenze nucleotidiche di individui caratterizzati dall'avere una

diversa storia clinica.

Il modello più ampiamente trattato in questa parte sarà il network motif, perché rappresenta, nell'ambito delle rete biologiche, una struttura atta a rappresentare proprio le interconnessioni che si creano tra i vari componenti cellulari, per poi ricercare, tra queste, degli schemi relazionali ricorrenti.

Per quanto riguarda il terzo capitolo, invece, si occuperà di esplicare in cosa consiste il Data Mining ed i metodi principali che sono stati ideati per estrarre informazioni utili, da un'ingente mole di dati.

È a questo proprosito che verrà introdotta la definizione di Pattern Mining e di cosa significhi cercare di identificare un modello o un sottografo che si presenti, mantenendo la stessa struttura costitutiva, all'interno della rete biologica di interesse.

Nel quarto capitolo verrà presentato l'algoritmo impiegato per effettuare delle analisi su quattro datasets di campioni, distinti in due gruppi, sulla base dello stato clinico dei pazienti esaminati.

Dopo aver presentato la tecnica su cui si basa l'algoritmo implementato nel software ed il modello di rete che è stato adottato per rappresentare i due differenti subsets di campioni, che consistono, rispettivamente, nell'insieme degli individui affetti da una specifica patologia e dall'insieme di quelli che, invece, non la manifestano, si definisce il problema che si vuole risolvere.

L'obiettivo che l'algoritmo si propone di raggiungere è, infatti, quello di ricercare, in ognuna delle due reti sopracitate, quali siano i patterns o sottografi eccezionali, ovvero quelli più discriminanti ed i nomi delle coppie di geni coinvolte in queste relazioni.

Segue, dunque, descrizione delle operazioni effettuate per rendere i dati idonei ad essere processati dal software e dell'elaborazione dei risultati ottenuti dopo l'analisi.

Infine, nel quinto capitolo, si è documento, anche tramite l'ausilio dei grafici, quello che si è ottenuto durante le due fasi di sperimentazione.

Quello che si vuole valutare, per ciascuno dei due grafi associati ad ogni set di campioni in esame, è come il numero di patterns che l'algoritmo si troverà ad analizzare, dipenda dal variare di alcuni parametri che incideranno sulle caratteristiche assunte dalle stesse reti.

Verrà mostrato, quindi, un tipo di approccio che si propone di identificare e di conseguenza, mettere in risalto, le differenze che si manifestano nei diversi profili genomici degli individui.

Il fine, allora, è quello di fornire uno strumento che permetta di risalire a quale siano i geni effettivamente coinvolti in queste strutture ricorrenti, classificate come discriminanti e tentare di utilizzare questa conoscenza, per prevedere ad esempio, eventuali disfunzioni o malattie che potrebbero manifestarsi nel paziente a cui appartiene il campione genetico analizzato.

Capitolo 1: Reti Biologiche

1.1 Definizione di rete biologica

Il termine "rete", in generale, si riferisce ad un insieme di linee, reali o ideali, che si intrecciano tra loro, formando incroci e nodi, dando luogo, così, ad una struttura complessa.

Esistono differenti tipi di reti e di modelli organizzativi, che possono essere classificati secondo diversi aspetti, sulla base dell'ambito che descrivono o della tipologia delle entità che compongono la struttura.
La maggior parte delle reti esistenti, siano esse sociali, biologiche o tecnologiche, è rappresentato da un modello basato sul concetto di rete, o più precisamente, di "rete complessa".

Una rete è detta complessa quando è costituita da un elevato numero di nodi e presenta alcune caratteristiche topologiche non immediatamente intuibili e rilevabili nelle reti più semplici, come:

- il consistente grado di distribuzione dei nodi;
- l'elevato coefficiente di raggruppamento;
- l'evidenza di una struttura gerarchica.

La qualità che accomuna tutti i tipi di reti complesse con cui è possibile modellare gli svariati fenomeni presenti in natura è l'elevata connettività, ovvero, presa una qualunque coppia di nodi dalla topologia della rete, tra questi è possibile costruire un percorso che li colleghi, composto da un numero esiguo di legami.[1]

Questa proprietà è correlata al concetto di “effetto del mondo piccolo”, conosciuto anche come “Teoria dei sei gradi di separazione”, ovvero: qualunque sia il grado di complessità di una rete, presi due individui qualsiasi, essi possono essere messi in collegamento attraverso un percorso costituito da pochi passaggi. Infatti, i nodi che compongono la rete, non rappresentano altro che delle persone tra le quali sussistono delle relazioni, espresse tramite degli archi e pari, in numero, alla lunghezza media di questi legami.

Tutto ciò è stato dimostrato dal matematico Duncan J.Watts per via algebrica [2], riprendendo gli studi sociali di M. Granovetter e ancor prima di S. Milgram ed andando quindi a costituire uno dei principali oggetti di interesse della disciplina della teoria delle reti.

Un particolare tipo di reti complesse è rappresentato dalle reti biologiche, un modello utilizzato per descrivere e comprendere i meccanismi e le dinamiche dei processi biologici ed il modo in cui gli organismi che ne sono parte interagiscano tra loro.

Più precisamente, questo tipo di reti è impiegato per descrivere le interazioni che sussistono tra le componenti cellulari di un organismo, quali proteine e geni, oppure per modellare le associazioni tra il suo fenotipo e genotipo.

1.2 Rappresentazione di una rete biologica

L'interesse nel dover rappresentare gli interattomi come reti biologiche nacque con la biologia dei sistemi, una disciplina orientata allo studio degli esseri viventi come un aggregato di macromolecole interagenti, che si evolvono nel tempo, ossia focalizzata sull'interazione dinamica delle parti di cui il sistema cellulare è composto.

Una qualsiasi rete cellulare, allora, dal punto di vista matematico, verrà rappresentata tramite un grafo non orientato, in cui i nodi rappresentano le entità biologiche coinvolte nel processo di interazione, come proteine, RNA molecolare e sequenze di geni e gli archi definiscono le associazioni che sussistono tra queste.

1.2.1 Definizione di grafo

Un grafo non orientato G = (V, E) è definito da un insieme finito V (G) = {v_1, . . . , v_n } di elementi, detti nodi o vertici e da un insieme E(G) = {e_1, . . . , e_m} di coppie non ordinate di nodi, dette archi.

Inoltre, dato l'arco *e = (v, w) = (w, v)*, allora i nodi *v* e *w* sono detti estremi di *e*, e si dice che l'arco *e* incida su *u* e *v*.

Questo può essere rappresentato tramite una lista di adiacenze, mantenendo un elenco dei nodi presenti nel grafo oppure utilizzando una matrice di adiacenza, che risulta essere più indicata per descrivere un grafo denso e che dispone di un elevato numero di archi tra i suoi nodi.

Nello specifico, dato un grafo (V,E) con |V| = n, la sua matrice di adiacenza è la matrice n×n il cui generico elemento a(i,j) è così definito:

- 1 se (i,j) $\in$ E ,
- 0 altrimenti,

dove N è il numero dei nodi del grafo ed ognuna delle sue celle (i,j) contiene un valore booleano pari a true, se il nodo i è collegato al nodo j.

Nel caso di grafo orientato la relazione di adiacenza è simmetrica, così come la matrice che lo descrive.

Un grafo è definito da alcune caratteristiche, quali:

- l'ordine, ovvero il numero di vertici |V|;
- la dimensione |E| , cioè il numero di archi che collegano i nodi;
- il grado di uno specifico vertice, che consiste nel numero di archi che incidono sullo stesso;

Dato un grafo non orientato G, si consideri come suo "grado massimo" quello del vertice con il maggior numero di archi incidenti e come suo "grado minimo" quello del vertice che, invece, è caratterizzato dall'aver il minor numero di archi incidenti.

1.2.2 Tipi di grafo

Sulla base di questi due parametri è possibile effettuare una distinzione tra due tipi di grafi, ovvero tra grafo regolare e grafo casuale:

- si definisce un grafo come regolare quando il grado massimo ed il grado minimo coincidono in numero e quindi assumono entrambi un valore pari a k;
- si definisce un grafo come casuale quando si specifica il suo grado come "grado medio", dato dal rapporto tra la somma di tutti i suoi gradi ed il numero complessivo dei nodi che lo compongono.

In conclusione, un grafo non orientato, costituito da N nodi, è detto "casuale" se gli n archi che definiscono le relazioni tra i suoi vertici sono stati scelti *casualmente* tra tutte le possibili combinazioni di collegamenti, ovvero tra gli

$\frac{N(N-1)}{2}$ archi che era possibile tracciare. [3]

La teoria dei grafi casuali si basa, quindi, sul principio che la probabilità di connessione sia uguale per qualsiasi coppia di nodi della rete e che questa si distribuisca in modo casuale.

Le reti complesse e casuali, allora, si instaurano scegliendo due nodi qualsiasi e collegandoli tra loro.

Se in seguito verranno aggiunti altri archi, in modo tale che su ogni nodo ne incida almeno uno, si otterrà un unico insieme di nodi interamente connesso.

Infatti, le reti casuali sono caratterizzate dall'essere strettamente connesse, dall'avere un basso diametro, definito come la distanza massima che sussiste tra ogni coppia di nodi, in termini di numero di link necessari al loro collegamento e dall'adottare un tipo di distribuzione binomiale, poichè la presenza o l'assenza di un arco risulta indipendente da quella degli altri archi.

Un modello intermedio tra la tipologia di rete casuale e quella regolare consiste nella rete a "piccolo mondo", definito da Watts e Strogatz, nel 1998.

Per definire una rete di questo tipo occorre partire da una rete circolare ordinata, in cui ogni nodo è provvisto di un legame con i quattro nodi ad esso più vicini. Si introducono, in seguito, alcune connessioni tra dei nodi scelti a caso, secondo una probabilità p.

Al variare di questo parametro p è possibile monitorare il passaggio da una rete regolare, caratterizzata da un valore di p pari a 0, ad una rete casuale, con una probabilità di p uguale a 1.

Per valori intermedi, invece, si ottiene la struttura di rete detta "piccolo-mondo".

[1]

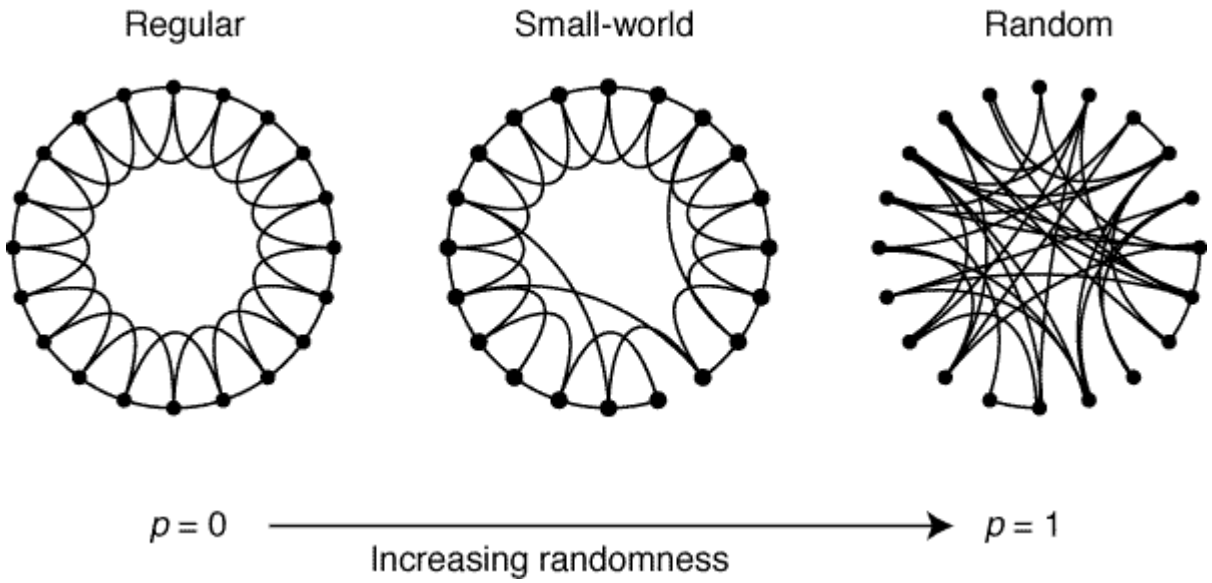


Immagine 1: Tre diversi tipi di rete, al variare della probabilità p.

Queste "small-world networks" sono caratterizzate dall'avere un basso grado di separazione, poichè ogni coppia di nodi comunica attraverso un numero minimo di collegamenti e da un elevato coefficiente di clustering, che comporta la presenza di esigui gruppi di nodi altamente connessi al loro interno e collegati, inoltre, ad altri cluster tramite legami più deboli.

1.2.3 Rete reale

Le reti reali rappresentano il mezzo di cui la teoria dei sistemi complessi dispone per descrivere i fenomeni della natura.

Con il termine "sistema complesso" si intende un sistema fisico definito dalla presenza di numerosi elementi che interagiscono in maniera non lineare e che sono fortemente interdipendenti gli uni con gli altri, caratterizzato da proprietà,

dette emergenti, quali la capacità di auto-organizzazione.[4]

Il modello più idoneo per descrivere i sistemi complessi è proprio quello delle reti a piccolo mondo, come si evince anche dalla topologia della rete delle reazioni biochimiche che governa il metabolismo dell'Escherichia Coli, studiata da Fell e Wagner.[5][6]

Tuttavia, negli anni, questo modello si è rilevato insufficiente perchè adatto a descrivere soltato reti egualitarie, ovvero un tipo di reti in cui tutti i nodi sono uguali.

Inoltre, una rete small-world presenta una distribuzione di tipo poissoniano, mentre, in natura, una rete reale presenta una distribuzione di gradi totalmente diversa.

Infatti, si è scoperto che alcune reti reali possiedono nodi altamente connessi, detti hubs, che fungono il ruolo di centri, favorendo così la rapidità della trasmissione dell'informazione e ostacolando, invece, la frammentazione della rete.

Questi hubs sono dei nodi che dispongono di un numero elevato di collegamenti che li legano l'un l'altro, rispetto ad altri nodi comuni che costituiscono la rete.

A causa di questo tipo particolare di nodi, allora, la rete assume una struttura gerarchica, anzichè egualitaria, poichè i nodi che la compongono non rivestono tutti lo stesso ruolo.

Questa caratteristica è stata evidenziata nel modello di rete ad invarianza di scala, elaborato da Barabasi ed Albert.[7]

Inoltre, la distribuzione dei gradi dei nodi non osserva più il modello

poissoniano ma la legge di potenza, in quanto un nodo con un ingente numero di collegamenti ha più probabilità di acquisirne di altri, nel momento in cui si inseriscono nella rete nuovi nodi.

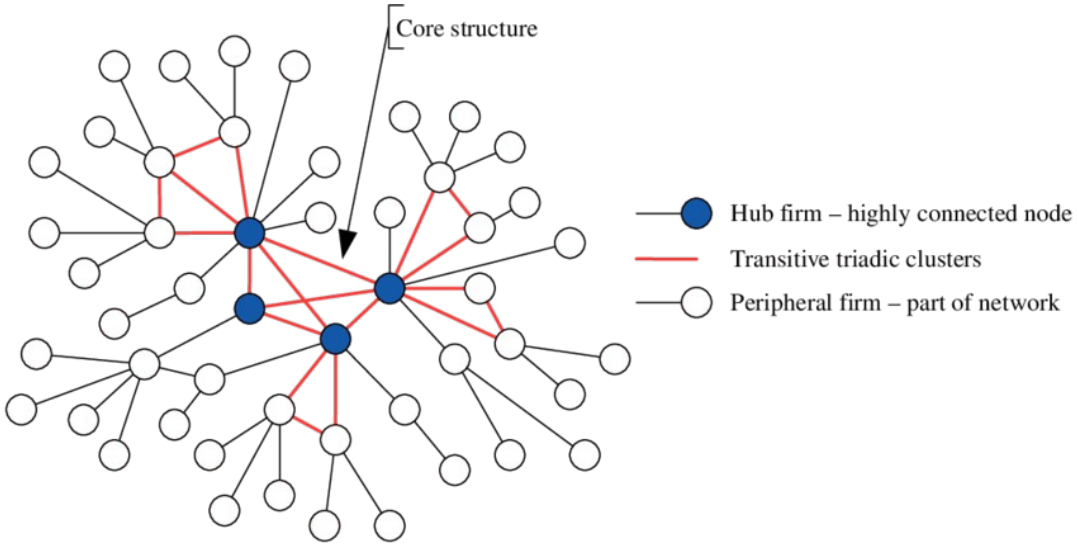


Immagine 2: Struttura di una rete ad invarianza di scala

In conclusione, una rete reale, è un tipo di rete complessa, caratterizzata da aspetti topologici non banali, capace di evolversi nel tempo e mutando, così, la sua conformazione è costituita da elementi che non posseggono in maniera esclusiva nè le peculiarità delle reti regolari, nè quelle delle reti casuali.

Un esempio esaustivo di rete reale è costituito dalle reti biologiche e più nello specifico dalle reti cellulari, ovvero quelle che modellano le interazioni che avvengono tra gli elementi che compongono la cellula, volte al soddisfacimento di specifiche funzioni biologiche.

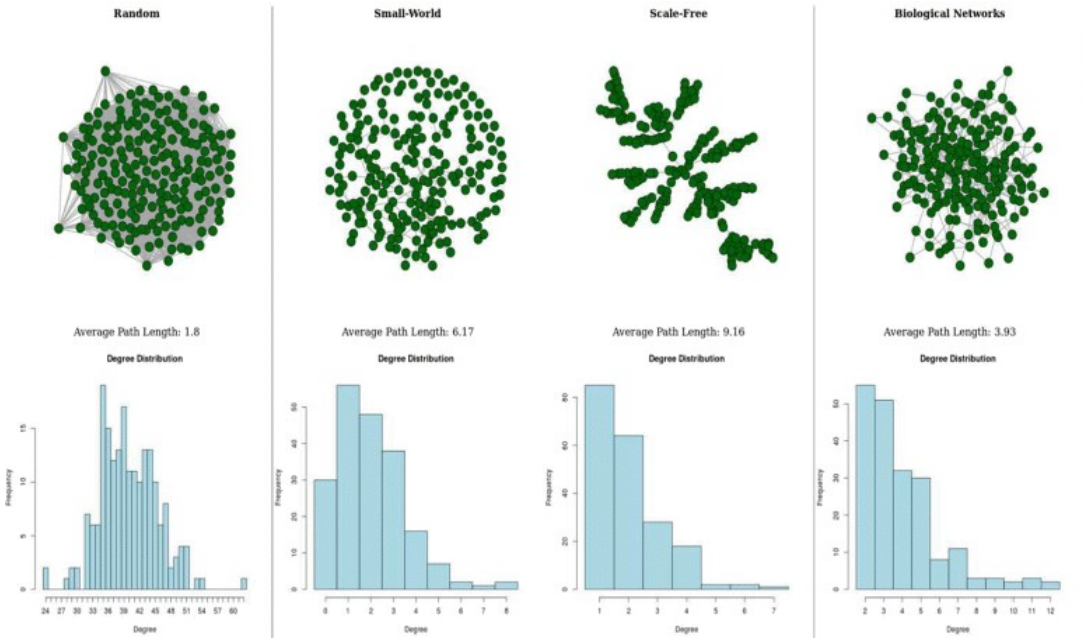


Immagine 3: Distribuzione dei gradi dei nodi nelle varie tipologie di rete

La complessità è dovuta al fatto che molte caratteristiche della cellula non sono riconducibili all'apporto da parte di una sola molecola ma, anzi, sono il risultato di interazioni complesse tra le componenti cellulari, quali proteine, DNA, RNA ed altre entità geniche.

È per questo che il tipo di approccio utilizzato dalle reti biologiche per modellare queste funzioni, come i meccanismi metabolici o di regolazione genica, non è più fondato su una visione "molecolare", basata sul considerare la cellula come totalità, ma che si focalizza sul considerarla come aggregato di moduli, composti ognuno da vari tipi di molecole, che interagiscono tra loro. [8]

1.3 Tipologie di reti biologiche

All'interno della cellula possono essere identificate diverse reti:

• le reti metaboliche, i cui nodi sono i composti chimici ed i links le reazioni;

• le reti trascrizionali o di regolazione genica, i cui nodi sono costituiti da geni e proteine ed i rispettivi collegamenti sono rappresentati dalle interazioni biochimiche tra questi;

• le reti di interazione tra proteine;

• le reti di interazione genica;

• le reti di segnalazione intracellulare;

Tutte queste reti non sono indipendenti, ma fanno parte di un'unica rete cellulare, comprendente tutte le componenti della cellula connesse, a loro volta, da tutte le interazioni fisiologicamente rilevanti, da quelle biochimiche a quelle fisiche.

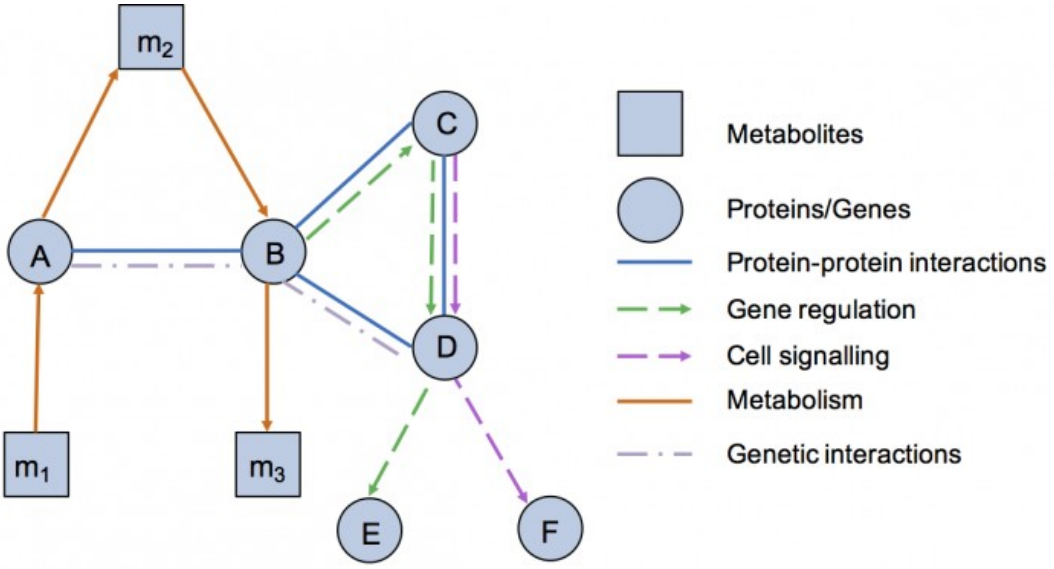


Immagine 4: Struttura di una rete cellulare

Segue una descrizione di alcuni tipi di reti biologiche, rilevanti per il prosieguo di questo studio, distinte tra loro in base allo scenario di applicazione ed agli elementi coinvolti nell'insieme complessivo di interazioni molecolari che avvengono in una qualunque cellula, definito come "interattoma".

Questo termine fu coniato nel 1999 da un gruppo di scienziati francesi ed in genere, con il concetto di interattoma, ci si riferisce alle interazioni che avvengono tra le molecole.

Queste possono essere di tipo diretto, se si verifica un'interazione fisica tra le molecole coinvolte, come nel caso di interazioni tra proteine, presenti sulla membrana cellulare di due cellule vicine, oppure di tipo indiretto, se la comunicazione avviene, ad esempio, tramite la secrezione di sostanze, ma comunque senza contiguità o compenetrazione fisica, come nel caso dell'interazione genica.

Allora, vista l'intricatezza e la complessità di questi legami, dal punto di vista formale, la rete che descrive l'interattoma di un organismo deve essere modellata secondo un approccio che sia il più possibile flessibile e ricco di informazioni biologiche aggiuntive sul tipo di collegamenti che si instaurano tra geni e proteine, tenendo conto anche del rumore di cui sono affetti i dati da utilizzare [9], proprio a causa della sovrabbondanza di interazioni molecolari, che, inoltre, cambiano nel tempo, in risposta a stimoli esterni, ma anche intracellulari.

1.3.1 Rete di interazione tra proteine

Le reti di interazione tra proteine consistono nella rappresentazione delle associazioni di complessi proteici formati da eventi biochimici, detti "pathways" e da forze elettrostatiche che rappresentano una specifica funzione biologica,

dunque, sono fondamentali per ogni processo che si svolge nella cellula.

Dal punto di vista formale, questo tipo di rete viene modellato tramite un grafo non orientato, denominato "Protein-Protein Interaction network", in cui i nodi rappresentano le proteine coinvolte nell'interazione e gli archi le loro interazioni.

I nodi possono essere etichettati con i nomi propri delle proteine oppure tramite dei codici identificativi. Per quanto riguarda gli archi, invece, saranno caratterizzati da pesi che rapprensentano il valore dell'affidabilità o "reliability"[10] specifica per quella particolare interazione, ottenuto tramite varie tecniche sperimentali e differenti approcci, volti proprio a trovare un metodo di misurazione accurato [11].

Inoltre, per quanto riguarda l'aspetto visuale, ogni nodo della rete verrà contraddistindo da un colore, che sarà indice dell'effetto fenotipico che comporterebbe la rimozione della proteina che identifica.

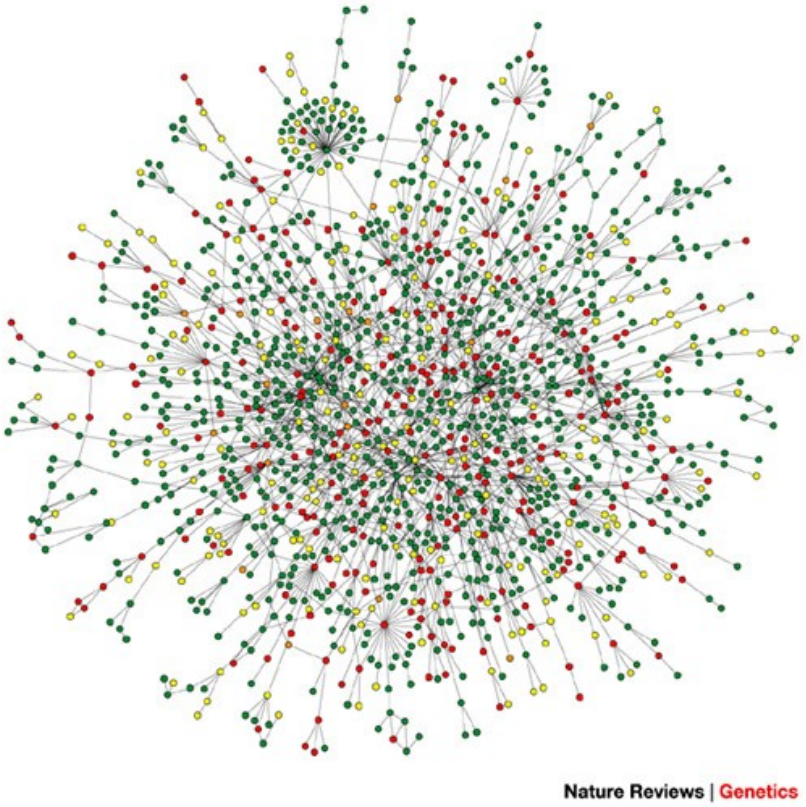


Immagine 5: Esempio di Protein Protein Interaction Network dell'organismo unicellulare Saccharomyces cerevisiae.

Nell'immagine riportata a titolo esemplificativo, il colore che contraddistingue ogni proteina indica che risulta essere caratterizzata dal manifestare una crescita lenta se arancione, oppure che la rimozione della suddetta sarebbe letale, se è colorata di rosso, non letale, se verde, oppure che avrebbe un effetto tuttora sconosciuto se è identificata con il colore giallo. [12]

1.3.2 Reti di interazione tra geni

Le reti di interazioni geniche sono utilizzate nella rappresentazione del rapporto funzionale che sussiste tra i vari geni che sono coinvolti, descrivendo il modo in cui questa interazione avviene.

In genere, dal punto di vista formale, i geni sono identificati come i nodi del grafo orientato che li rappresenta, mentre gli archi esemplificano le loro relazioni, specificando il ruolo che ricopre ogni gene nell'interazione di interesse, in base alla direzione assunta dell'arco che lo collega all'altro gene coinvolto.

Principalmente, queste reti si utilizzano per modellare le interazioni genetiche che condizionano la relazione che sussiste tra il fenotipo ed il genotipo di un organismo, dovuto proprio al rapporto funzionale che si instaura tra due o più geni, oppure quelle che riguardano il modo in cui si manifesta l'espressione genica.

L'impiego di questo tipo di approccio si è reso necessario dal momento che la ricerca si è indirizzata sul sequenziamento del genoma umano, come dimostra lo sviluppo del "Human Genome Project" o "progetto del genoma umano", una ricerca scientifica di ambito internazionale, volta a perseguire il sequenziamento del DNA, ovvero la determinazione dell'ordine dei diversi nucleotidi, cioè delle quattro basi azotate che li differenziano, quali Adenina,

Citosina, Guanina e Timina, che costituiscono l'acido nucleico.

Siccome la sequenza del DNA contiene tutte le informazioni genetiche ereditarie che sono alla base dello sviluppo degli organismi viventi, poiché al suo interno vengono codificati i geni ed i meccanismi che regolano proprio l'espressione genica, ovvero il processo tramite il quale le informazioni genetiche vengono convertite in macromolecole funzionali, come le proteine, la conoscenza del genoma può essere sfruttata soprattutto in ambito medico.

Infatti, è utilizzata per identificare e diagnosticare malattie ereditarie e sviluppare trattamenti innovativi, più specifici perché indirizzati ad agire, in modo mirato, sulla mutazione o su una qualsiasi entità genica che provochi la patologia di interesse, tenendo conto anche delle caratteristiche cliniche del soggetto che deve essere sottoposto alla cura, e di conseguenza più efficaci, oltre che per la produzione di medicinali che argino il diffondersi di malattie contagiose, elaborati ad-hoc sulla base della struttura genetica degli agenti patogeni che le causano.

Il contributo dell'utilizzo di queste tecniche basate su reti sarebbe notevole anche per quanto riguarda la comprensione delle differenze che che si riscontrano in pazienti affetti dalla stessa patologia, soprattutto nei casi di malattie tumorali, poiché un approccio orientato al discernimento dei diversi profili genomici, atto ad identificare se i geni coinvolti nella mutazione appartengano allo stesso "pathway" o percorso biologico soggetto ad alterazioni, e concorrano, quindi, al manifestarsi di malattie simili non soltanto in termini di genotipo, ma anche in termini di fenotipo.

L'obiettivo sarebbe quello di identificare le mutazioni "driver" del cancro, ovvero quelle mutazioni genetiche di tipo somatico che svolgono un ruolo primario nella genesi e nello sviluppo delle cellule tumorali.

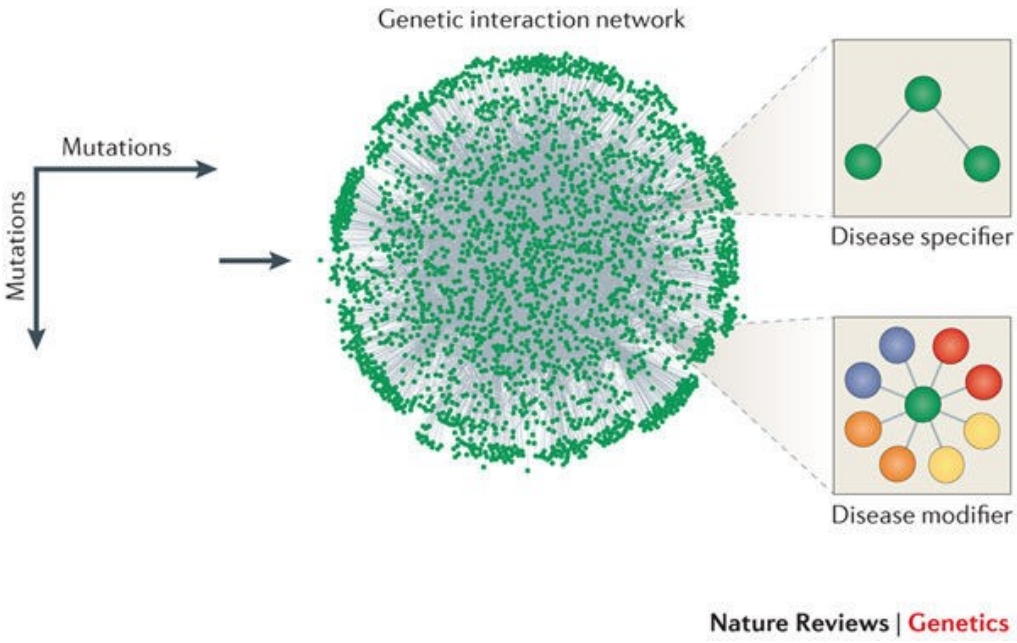


Immagine 6: Struttura di rete di interazione genica

Tra i tipi di reti geniche che sfruttano i dati sul genoma degli esseri viventi, si considerino le reti basate sulle associazioni genotipo-fenotipo e quelle focalizzate sulla coespressione dei geni di interesse.

1.3.3 Reti di associazione tra genotipo e fenotipo

Si definisce "Genotipo" la costituzione genetica di un organismo, ovvero il suo patrimonio ereditario.

Si definisce "Fenotipo" l'insieme delle caratteristiche morfologiche e funzionali di un qualunque essere vivente, determinate dall'interazione fra il suo genotipo e l'ambiente in cui vive.

Allora, questo descrive il modo in cui l'espressione genica si manifesta , ovvero come tutte le caratteristiche contenute nel DNA di un organismo vengano convertite in macromolecole funzionali, come le proteine, caratterizzando

proprio la sua struttura ed i suo compotamento.

Le "Genotype-Phenotype networks", quindi, si occupano di descrivere, la correlazione che sussite tra il fenotipo ed il genotipo di un individuo, qualora questa fosse presente, per specifiche coppie di geni, fornendo uno strumento di supporto ai vari metodi che si sono sviluppati proprio con l'obiettivo di tracciare questa associazione[13] [14], in quanto, gli effetti fenotipici che si riscontrano in individui affetti patologie, come i tumori, risultano essere conseguenza delle alterazioni genetiche, o mutazioni, che incorrono quando si verificano delle disfunzioni nelle attività biologiche di una cellula.[15]

Una rete che descrive questa correlazione tra fenotipo e genotipo, creata per l'analisi di una specifica patologia, può essere rappresentata come un grafo bipartito, cioè un grafo in cui l'insieme dei suoi vertici può essere partizionato in due sottoinsiemi, tali che ogni vertice di uno dei due partizionamenti sia collegato solamente ai vertici dell'altro.

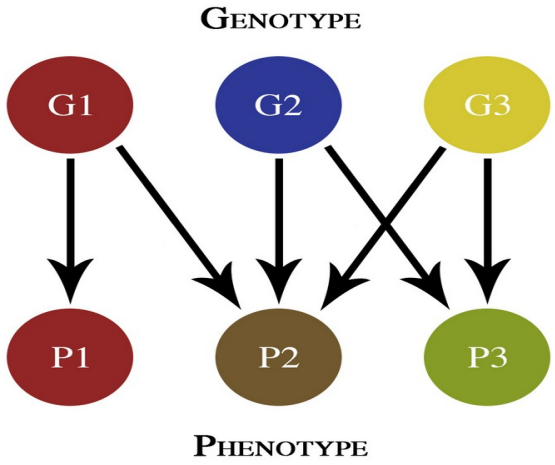


Immagine 7: Grafo bipartito tra genotipo e fenotipo

Nel caso in cui si volesse rappresentare il grafo bipartito di un insieme definito di patologie, per tracciare la predisposizione che ha un individuo di essere affetto da uno specifico disturbo, noto il suo patrimonio genetico, questo

approccio condurrebbe alla formazione di due sotto-grafi [17], quali:

- una prima rete, in cui le malattie andranno a costituire i nodi, collegati tra loro tramite gli archi del grafo soltanto se accomunati dalla presenza di un determinato gene;
- una seconda rete, in cui i nodi saranno i geni, associati tra loro tramite gli archi, se entrambi fanno parte del genoma di un individuo affetto da quello specifico disturbo.

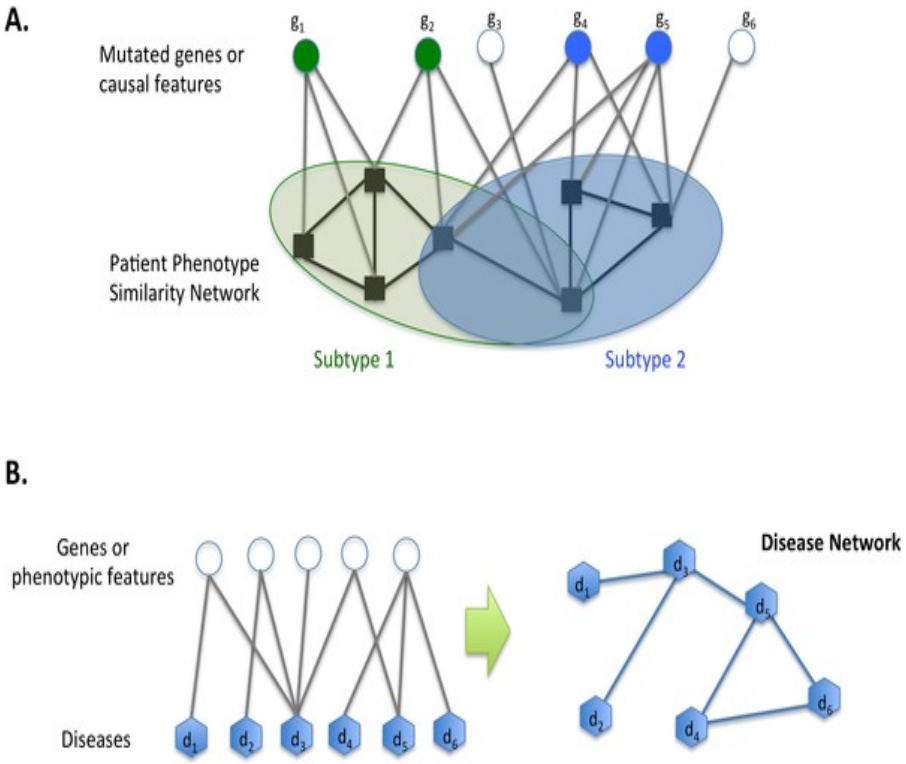


Immagine 8: Rappresentazione di una Disease Network tramite grafo bipartito

1.3.4 Reti di co-espressione genica

Con "espressione genica" si intende il processo di formazione dell'RNA messaggero, detto mRNA, sul filamento stampo di DNA e la sintesi della proteina da parte di uno specifico mRNA sui ribosmi del citoplasma.

Questo è soggetto ad una serie di meccanismi di regolazione genica, che permettono, ad una cellula, di esprimere soltanto alcuni dei geni presenti nel suo corredo genetico. Tuttavia, i geni che non sono stati manifestati nelle cellule differenziate non verranno distrutti, o mutati, ma manterranno il loro potenziale di essere espressi.

Questo complesso insieme di interazioni, che controlla ed influenza l'espressione genica e, di conseguena, il comportamento cellulare di un organismo, è descritto da un tipo di reti dette "Genetic Regulation Network".

I dati su cui si strutturano queste reti, introdotte per la prima volta da Butte e Kohane con il nome di "reti di rilevanza"[18], sono basati su un raggruppamento di informazioni relative alla *co-espressione genica*, ovvero sulla similarità che è possibile identificare, tra due o più geni, relativamente al loro pattern di espressione.

Generalmente, ad esempio, due geni che sono funzionalmente correlati, ovvero che concorrono alla realizzazione della medesima azione biochimica all'interno della cellula, in risposta alla variazione delle condizioni ambientali, permettendole di regolare le proprie funzioni interne e gli scambi che intraprende con l'esterno, risultano essere anche coespressi.

Infatti, la capacità di prevedere con precisione la funzione genica basata sulla sequenza di uno specifico gene si basa proprio sull'indentificazione e sulla caratterizzazione della similarità, ovvero della somiglianza che si riscontra nella sequenza di espressione tra il gene di interesse e quella dei geni di cui è già noto che siano coinvolti in una determinata attività biochimica all'interno

della cellula.

Però, siccome il solo riscontro di queste similarità di espressione non è sufficiente per attuare un'analisi accurata, ci si focalizza sul processo evolutivo che ha condotto al manifestarsi di questa somiglianza tra due sequenze geniche[19], secondo un approccio già adottato nella biologia comparata ed evolutiva [20], una disciplina che utilizza la variazione naturale e la diversità tra entità, siano queste dei profili genetici oppure intere comunità, per comprendere i modelli che caratterizzano gli organismi ed il ruolo che questi ricoprono nell'ecosistema.

Per quanto riguarda la struttura di una rete di co-espressione genica, i nodi che la compongono sono rappresentati dai geni, a ciascuno dei quali è associato un valore di espressione ricavato da un insieme di campioni, costituiti dai profili genici di alcuni organismi.

L'arco tra due nodi e quindi tra due specifici geni, verrà creato soltanto se i punteggi o "score" del *valore di espressione* determinato per ciascuno dei due geni analizzati, fornito dai campioni utilizzati come dati per la costruzione del grafo, risultano, in qualche modo, correlati tra loro, ovvero se sussiste una corrispondenza tra i due profili genici.

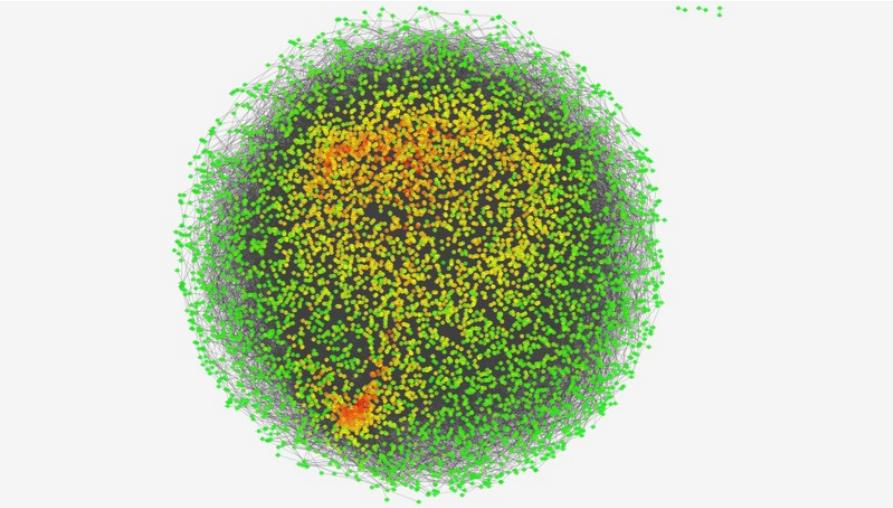


Immagine 9: Esempio di rete di co-espressione genica costruito sulla base dei valori di espressione genica di 7221 geni individuati in 18 pazienti affetti da tumore allo stomaco.

Nella costruzione di una generica rete di co-espressione genica è importante che vengano realizzati due passi principali, quali:

- la misurazione del valore di coespressione, da utilizzare per il calcolo del punteggio di somiglianza che interessa ogni coppia di geni della rete;
- la scelta della soglia di significatività o "threshold", in base a cui si eseguiranno dei tagli volti a selezionare soltanto le coppie di geni che hanno un punteggio di somiglianza superiore a questo valore , e perciò idonee ad essere considerate come significative e ad essere incluse nella rete, sottoforma di arco che collega i due geni interessati.

Siccome il valore di espressione di un gene può essere rappresentato utilizzando un vettore che contenga i valori effettivamente riscontrati con l'analisi dei diversi campioni, il calcolo del valore di coespressione tra due geni si riduce ad essere un'operazione che interessa due array numerici.

Questa può essere effettuata utilizzando differenti tipi di misurazione, tra i quali l'impiego del coefficiente di correlazione di Pearson risulta essere quello più utilizzato, nonostante presupponga che i dati che descrivono l'espressione genica seguano una distribuzione normale e che possa rilevare soltanto relazioni di tipo lineare, manifestando, quindi, degli svantaggi [21].

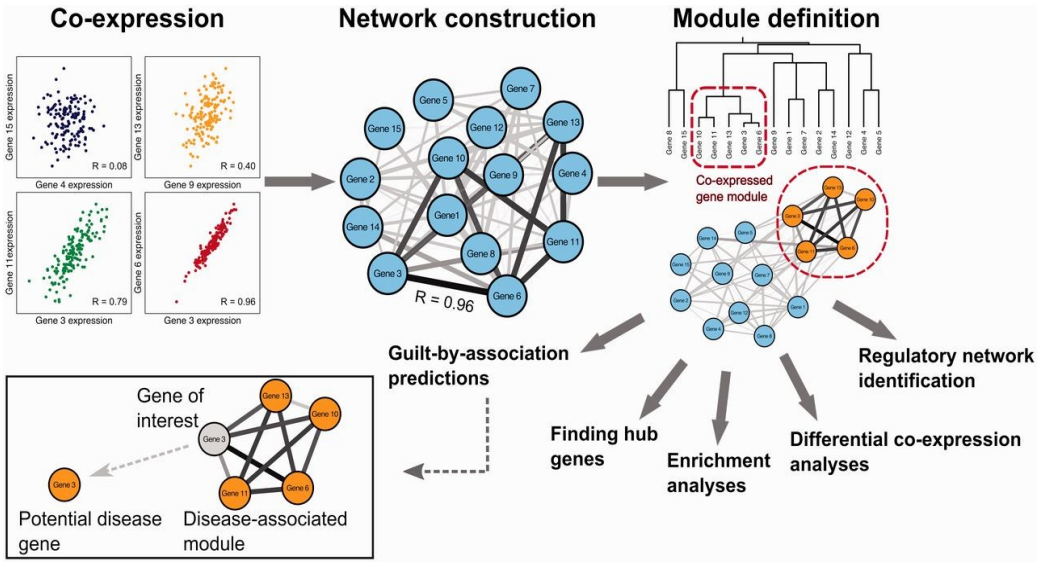


Immagine 10: Creazione di una rete di coespressione a partire dall'espressione genica

Sono numerose le ricerche che utilizzano le reti di coespressione genica per confrontare tra loro i singoli geni, integrando l'approccio classico, che si focalizza sull'identificare le funzioni che potrebbero essere influenzate dalla loro azione e sull'indagare sulla loro evoluzione storica [22], in una metodologia che tenga conto dell'interazione che si crea tra i diversi gruppi di geni e della possibilità che geni similari in termini di sequenza possano manifestarsi tramite differenze fenotipiche estremamente significative tra le varie specie.

A tale proposito, si riporta, come riscontro, lo studio effettuato sui geni coinvolti nel cervello umano e in quello degli scimpanzè, che ha evidenziato come, nonostante questi geni presentino una sequenza significativamente compatibile in termini di similarità, la loro manifestazione fenotipica sia, invece, notevolmente differente [23].

Allora, si rende necessario, ai fini di comprendere se i gruppi di geni che costituiscono queste similarità, ovvero i gruppi di geni coespressi, siano effettivamente coinvolti nell'adempimento delle stesse funzioni biologiche ed accumunati, di conseguenza, dalla stessa storia evolutiva, verificare se tra le coppie di geni di interesse esista un valore di correlazione sufficiente a dimostrare la loro coespressione, descrivendo poi una rete di coespressione genica che tracci queste corrispondenze, creando un arco tra i geni che risultano coespressi e la cui topologia possa essere analizzata, con lo scopo di raccogliere sempre più informazioni sulla biologia degli organismi [24].

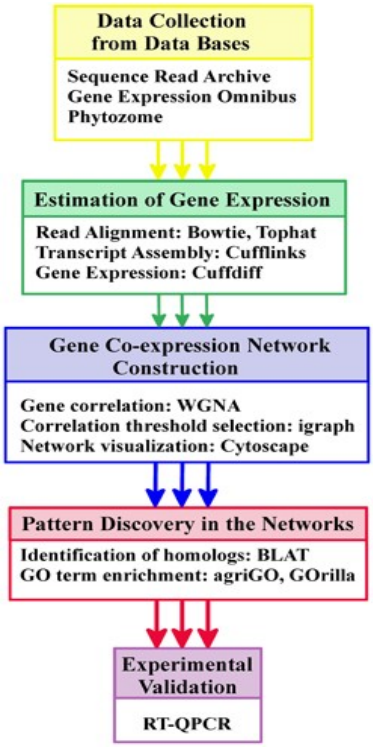


Immagine 11: Schema che descrive le fasi principali seguite per l'analisi delle reti di coespressione per Chlamydomonas, Physcomitrella e Arabidopsis, tre specie di piante.

1.4 Dati utilizzati

L'analisi dell'espressione genica ed il conseguente sviluppo di svariate metodologie atte ad identificare le sequenze geniche ed i relativi livelli di espressione, come le tecniche dei microarray, hanno agevolato e contribuito all'archiviazione e alla diffusione di un'ingente quantità di dati, relativa ai profili genomici esaminati nel corso degli anni.

Si consideri, ad esempio, proprio la tecnica che impiega la creazione di un microarray di DNA, che consiste in un gene chip utilizzato per analizzare, contemporaneamente, un insieme elevato di geni presenti all'interno di un campione del genoma o del trascrittoma di un organismo, in modo da poter confrontare le sequenze di espressione genica di un individuo sano con quelle di un individuo affetto, invece, dalla patologia di interesse, in modo da identificare quali geni siano coinvolti nell'insorgere di quest'ultima.

L'utilizzo di questo approccio, così come delle tecniche di nuova generazione, ha fatto sì che si creassero numerose banche di dati, atte ad immagazzinare dati biologici di differente natura, siano questi tassonomici, genomici, relativi ad entità virali o, più in generale, alle componenti cellulari, quali proteine ed acidi nucleici.

Tra le banche che contengono un elevato quantitativo di informazioni riguardanti, soprattutto, le macromiolecole biologiche, si riportano le seguenti:

- "European Nucleotide Archive" (ENA), un database che fornisce informazioni basate sul sequenziamento dei nucleotidi, mantenuto dal

  Laboratorio Europeo di Biologia Molecolare (EMBL), un istituto di ricerca sostenuto da venti paesi europei e dall'Australia, in qualità di membro

associato.

- "GenBank", mette a disposizione una raccolta di sequenze genetiche sia nucleotidiche che proteiche, complete di annotazioni funzionali e bibliografiche, estratte da molteplici pubblicazioni in campo medico. Questa banca dati, così come PubMed, è sostenuta dal Centro Nazionale per le Informazioni Biotecnologiche (NCBI), che dipende, a sua volta, dal National Institutes of Health (NIH) degli Stati Uniti.

- "DNA Data Bank of Japan" (DDBJ), una banca di dati biologici che si occupa di raccogliere le sequenze di DNA, istituita nel 1986 presso il National Insistute of Genetics (NIG), in Giappone.

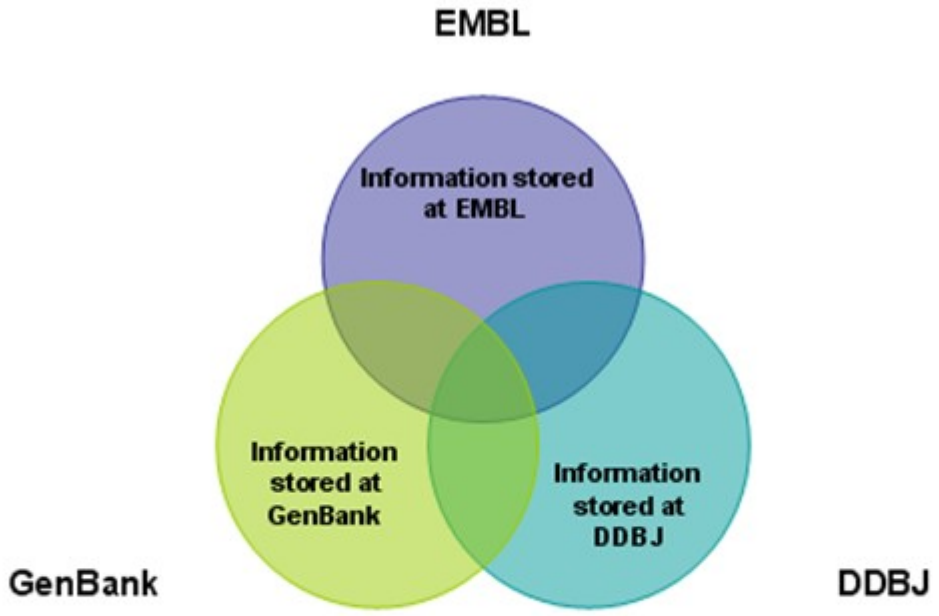


Immagine 12: Le informazioni raccolte da EMBL, GenBank e DDJB confluiscono in un'unica banca dati, aggiornata e condivisa.

Le tre banche dati sopraelencate hanno instaurato, tra loro, una stretta collaborazione, che favorisce lo scambio di informazioni ed il conseguente arricchimento ed aggiornamento giornaliero dei rispettivi database,

ufficializzata nella International Nucleotide Sequence Database Collaboration (INSDC), un'iniziativa che realizza una banca dati collettiva e condivisa tra i membri effettivi del comitato che la rappresenta, ovvero gli organi stessi che contribuiscono al deposito dei dati, corredati da informazioni sperimentali ed annotazioni sulle caratteristiche dei campioni da cui sono stati prelevati.

Secondo la politica INSDC, i dati raccolti sulle sequenze nucleotidiche sono liberamente accessibili e consultabili, tanto da non essere soggetti ad alcuna restrizione relativa al loro utilizzo, in qualsivoglia attività di ricerca e sperimentazione, di analisi e pubblicazione.

Inoltre, il supporto fornito da questa istituzione riguardo lo scambio delle informazioni relative a campioni di DNA e RNA estratti ed esaminati dagli scienziati di tutto il mondo, anche attraverso lo sviluppo e la manutenzione di strumenti che mette a loro disposizione, ha portato alla definizione di uno standard bioinformatico universalmente riconosciuto, che interessa il formato che i dati devono adottare per essere presentati, scambiati ed analizzati, oltre che utilizzati per lo sviluppo di nuovi sistemi e tecnologie.

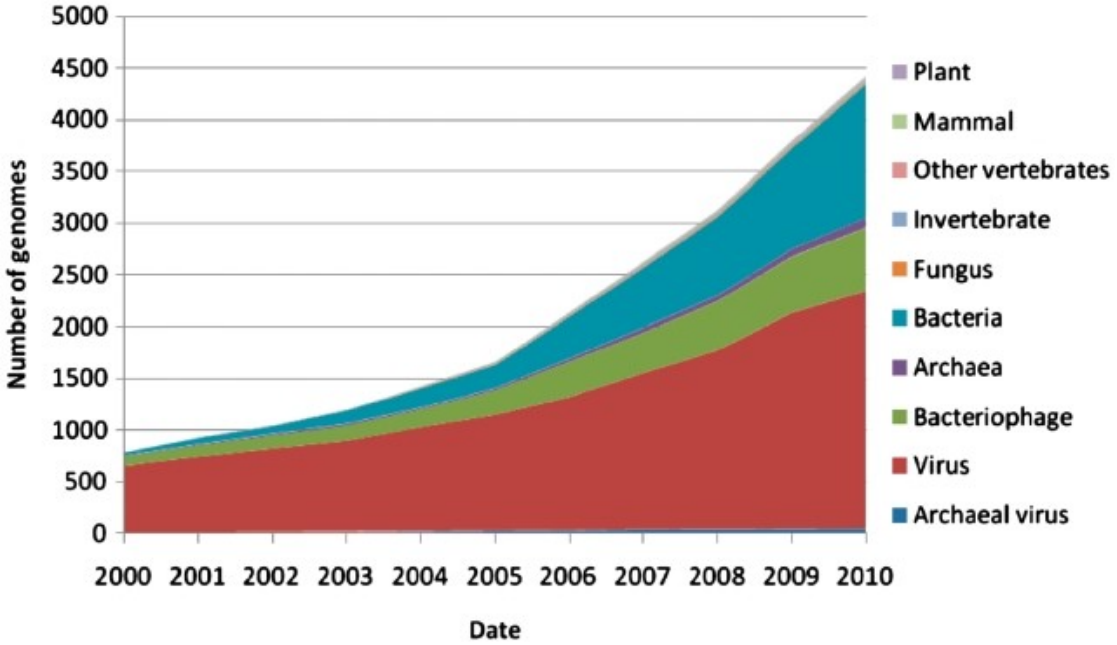


Immagine 13: Tipologia di dati raccolti dall'International Nucleotide Sequence Database Collaboration.

Capitolo 2: Tecniche di analisi

Dopo aver formalizzato i dati raccolti sottoforma di grafi, utilizzando il modello di rappresentazione descritto per le reti biologiche, sopraggiunge la necessità di definire un metodo che analizzi, agevolmente ed in maniera automatizzata, questa ingente quantità di informazioni, ed estragga, quindi, nuova conoscenza, esaminando la rete di interesse, sulla base della sua topologia, che evidenzia le relazioni che sussistono tra i nodi che la compongono, e del suo stato, che descrive, invece, le proprietà di questi e degli archi che li collegano.

L'obiettivo risulta essere, perciò, quello di riuscire a comparare, tra loro, reti biologiche di natura e dimensioni diverse, costituite da geni e proteine che sono composti da differenti sequenze di amminoacidi, il tutto per determinare, quali tra queste interazioni fisiche che avvengono tra le molecole coinvolte in un'analoga attività biochimica, possano essere rilevanti per la comprensione di uno specifico processo o delle cause che concorrono all'insorgere di una patologia, ed identificare, di conseguenza, gli elementi coinvolti e se questi svolgano, o meno, un ruolo cruciale, in base alla loro funzione.

Questo comporta, però, il dover confrontare dei grafi sparsi, che possono arrivare a raggiungere dimensioni dell'ordine dei 10.000 nodi e 40.000 archi, e risolvere, quindi, un problema di complessità notevole.

Nel dettaglio, dati due grafi G e H, si vuole determinare se in G sia possibile identificare un sottografo che risulti essere isomorfico al grafo H, dove per isomorfismo fra due grafi si intende che sussista una corrispondenza biunivoca tra i due insiemi di vertici.

Allora, la ricerca di un sottografo isomorfico risulta essere un problema decisionale di tipo NP-completo, ovvero non deterministico in tempo polinomiale.[25]

Segue un elenco delle principali tecniche di network comparison, ovvero una descrizione delle tecniche che sono state elaborate proprio per confrontare due o più reti di interazione genica, che possono raffigurare interazioni di diversa natura, ma anche tra organismi appartenenti a specie diverse.

2.1 Network alignment

L'analisi della rete incentrata sull'allineamento delle sequenze nucleotidiche si fonda sul mettere a confronto i due grafi che descrivono l'interazione genica di interesse, identificando regioni di similarità, ovvero andando ad individuare una sottorete, una struttura che si sia conservata, cioè che si sia mantenuta sostanzialmente invariata in entrambe le specie, in modo tale da essere considerata un vero e proprio modulo funzionale, sia questa un singolo percorso biologico dedito allo svolgimento di una specifica funzione cellulare, oppure un cluster di interazioni, che coinvolga, quindi, interi complessi proteici.

Questa tecnica di analisi può essere di diversi tipi, ovvero orientata ad una comparazione tra coppie di sequenze (pairwise alignment) oppure basata su un confronto multiplo (multiple alignment). Un'ulteriore distinzione può essere effettuata sull'approccio che si utilizza in base il criterio di "similarità" che si è definito, considerando, quindi, un allineamento di tipo locale oppure globale.

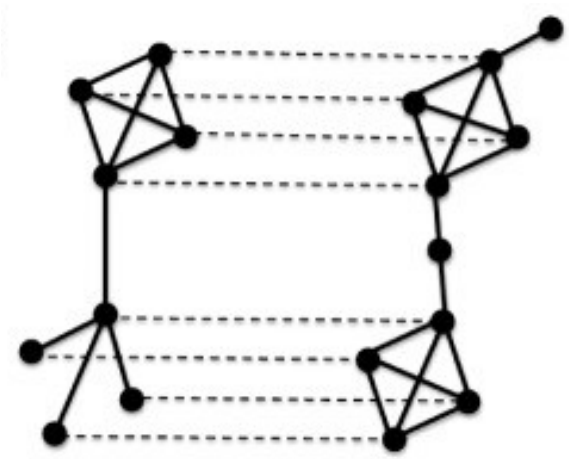


Immagine 14: Esempio di Global Alignment.

Nel caso di allineamento globale, si punta alla ricerca di un'unica regione isomorfica, che colleghi, secondo una corrispondenza uno-ad-uno, i nodi della prima rete con quelli della seconda.

Questo perchè si vuole misurare la similarità complessiva che interessa le reti analizzate, allineando ciascun nodo della rete più piccola, in termini di dimensione, con uno ed un solo nodo della rete più estesa.

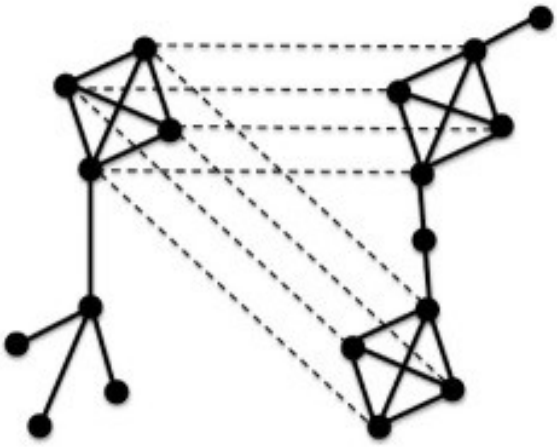


Immagine 15: Esempio di Local Alignment.

Per quanto riguarda la tecnica di allineamento locale, invece, è orientata alla ricerca di molteplici sottografi isomorfici tra le reti messe a confronto, di cui

ogni regione isomorfica che è stata identificata risulta essere indipendente dal modo in cui sono state costruite tutte le altre e caratterizzata da una corrispondenza di tipo molti-a-molti, poichè è consentita la presenza di accoppiamenti differenti per uno stesso nodo.

In conclusione, l'allineamento di sequenza è orientato alla predizione di funzioni e proprietà che interessano molecole biologiche, quali geni e proteine, con il fine di determinare nuovi tipi di interazioni e collegamenti che concorrono allo svolgimento dei vari processi cellulari, poichè l'identificazione di una subnetwrok che si è conservata nel corredo genetico di organismi appartenenti a specie differenti, costituita da un insieme di proteine che cooperano nell'attuazione della medesima funzione, permetterebbe di stabilire che anche le proteine rimanenti, non incluse nella suddetta sottorete, siano accomunate dal dover adempiere alla stessa funzionalità.

2.2 Network integration

La tecnica di network integration consiste nel combinare più reti geniche di tipo differente, costituite però da geni e proteine che appartengano, comunque, ad organismi della stessa specie ed al medesimo insieme di elementi, ovvero che i campioni di proteine o geni utilizzati per la costruzione di queste reti di interazione rimangano invariati, in qualità di nodi che le compongono.

Siccome ogni rete biologica descrive un aspetto di uno specifico processo cellulare, l'azione di esaminare un'unica network che includa, quindi, altre reti, ciascuna delle quali rappresenti una diversa interazione genica o proteica, permette di combinare dati che provengono da più campioni, in modo tale che vengano analizzati nel loro complesso.

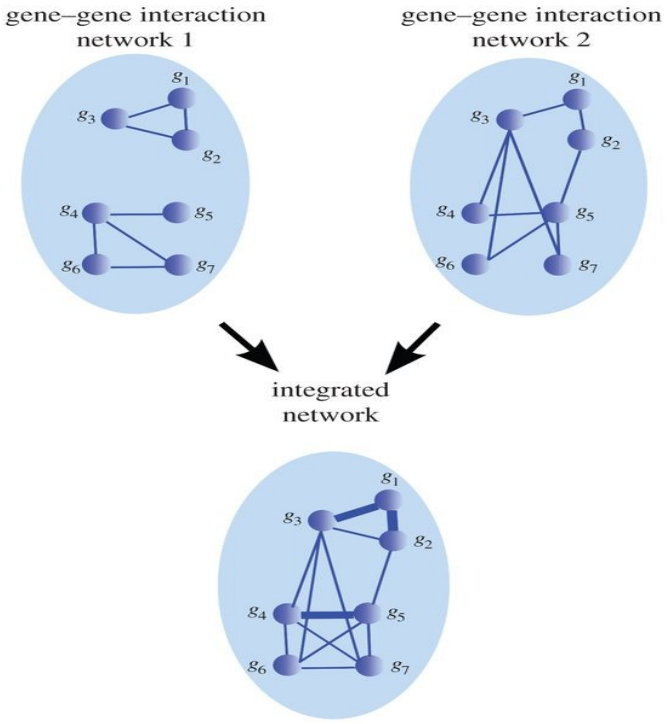


Immagine 16: Esempio di rete integrata, costituita da due reti di interazione genica.

In questo modo è possibile valutare, più agevolmente, quali tra i moduli funzionali che si instaurano, ovvero gli aggregati di proteine e geni che si mantengono inalterati nei vari profili genomici considerati, ricoprano un ruolo cruciale nello svolgimento di una specifica attività biochimica, e predirre, inoltre, in quale altre funzioni questi possono essere coinvolti.

2.3 Network querying

Il network querying consiste nell'analizzare una rete definita "target network", ai fini di identificare, al suo interno, tutte le occorrenze, determinate per similarità, di una sottorete di interesse, detta "query network", della quale è noto, a priori, che rappresenti un modulo funzionale.

Questo approccio permetterebbe di evidenziare la presenza di questi moduli

funzionali, ossia delle entità cellulari che risultano essere coinvolte, costantemente, in processi biologici espletati per mezzo di funzioni analoghe o quantomeno affini, anche in reti di cui non si conoscono le proprietà che le caratterizzano, proprio per mezzo di un confronto diretto.

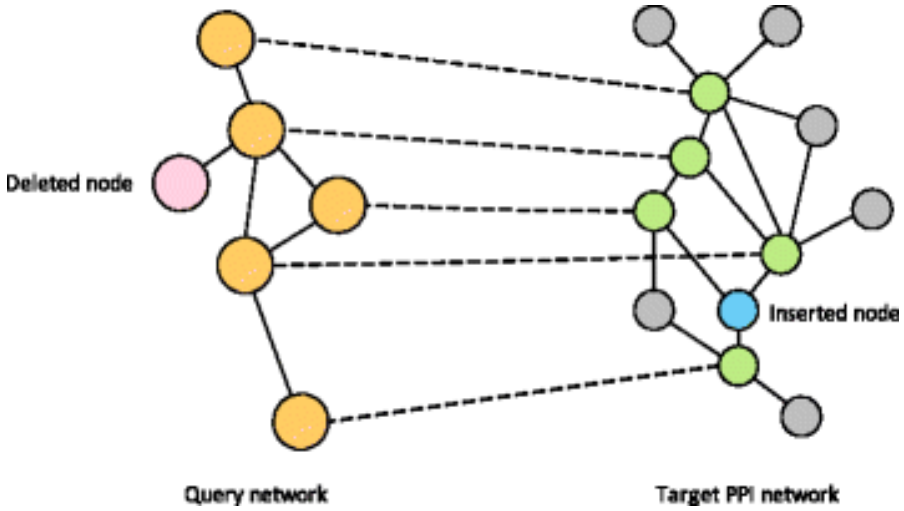


Immagine 17: Esempio di network querying, che evidenzia, nella rete target, la regione che corrisponde alla rete query.

2.4 Network clustering

Un altro metodo che è possibile sfruttare per l'elaborazione delle reti geniche è il network clustering, che consiste nel raggruppare tutti i geni e le proteine in degli agglomerati di nodi, definiti clusters, caratterizzati dall'aver attuato, tra loro, un elevato numero di interazioni.

Il criterio con cui viene effettuato il raggruppamento è, anche in questo caso, basato sul concetto di similarità, valutata tra coppie di geni o proteine.

Il risultato che si vuole ottenere è quello di chiarire e comprendere la struttura di queste reti di interazione, cercando di identificare le funzioni biologiche svolte dai membri dei clusters, cioè da tutte quelle entità cellulari il cui ruolo risulta essere ancora sconosciuto.

Generalmente, gli approcci per l'analisi delle reti biologiche che utilizzano la tecnica del network clustering sono basati sulla distanza, oppure su grafo [26].

La differenza consiste nella scelta delle caratteristiche che devono essere elaborate ai fini di effettuare la ricerca dei complessi proteici di interesse, perciò, nel caso di un approccio basato sulla distanza, il parametro da valutare ai fini di effettuare la clusterizzazione sarà proprio la distanza, ad esempio in termini di sequenza amminoacidica, che sussiste tra due geni o proteine.

Per quanto riguarda, invece, il clustering basato su grafo, le informazioni di rilievo risultano essere tutte correlate alla topologia che presenta la rete da analizzare, poichè si mira ad individuare, in essa, dei sottografi più densi, i cui nodi siano altamente connessi.

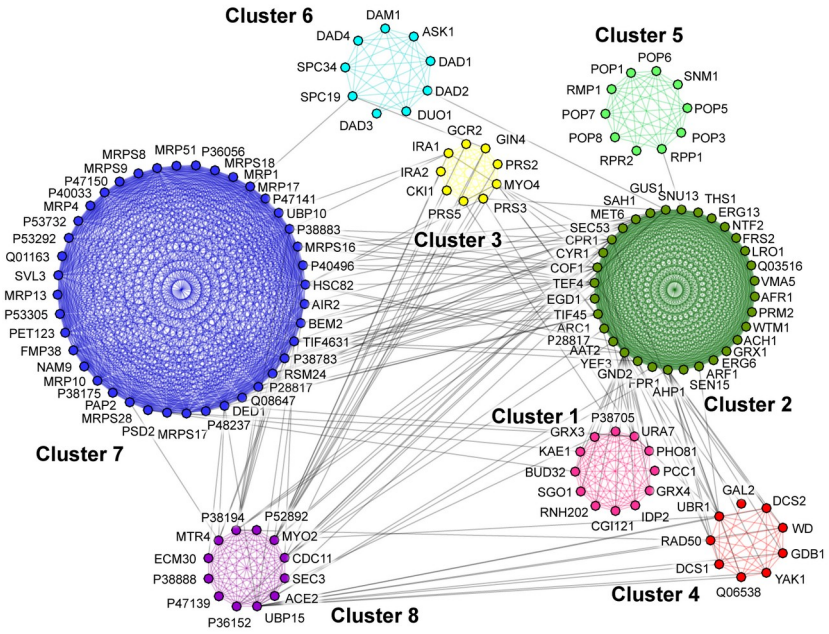


Immagine 18: Esempio di network clustering per l'identificazione e la predizione delle funzioni dei diversi complessi proteici.

2.5 Network motif extraction

Per definire l'approccio di network motif extraction occorre soffermarsi sul concetto di motif ed il significato che acquisisce se correlato al concetto di grafo.

Il termine "motif" o "motivo" è utilizzato per caratterizzare un oggetto che si ripete, frequentemente, nell'insieme di elementi a cui appartiene.

Considerando questa definizione come una proprietà che può essere manifestata in un grafo, si ottiene, quindi, l'espressione di "network motif", o "motivo di rete".

2.5.1 Definizione di network motif

Secondo la definizione di S. Shen-Orr, un "network motif" rappresenta dei patterns di interconnessioni che ricorrono, nella struttura rete analizzata, con una frequenza più elevata di quelli che possono essere, invece, riscontrati in una rete di tipo casuale [27].

Allora, data una rete biologica, si descrive come network mosfit una sottorete ricorrente e significativa dal punto di vista statisco [28], la cui struttura può essere riscontrata, quindi, in una qualsiasi parte della rete a cui appartiene ma anche in reti differenti.

Diversamente dai moduli funzionali, cioè i complessi proteici che costituiscono l'oggetto della ricerca di altre tecniche di analisi delle reti biologiche, descritti, in precedenza, come dei sottografi indipendenti, dal punto di vista funzionale, dagli altri nodi della rete, poichè ciascuno di questi agglomerati svolge una specifica attività cellulare, questi "motivi di rete" sono da considerarsi come dei

sottografi isomorfi costituiti da una struttura piccola e definita, che non necessitano di alcuna autonomia per poter manifestarsi, comunque, con rilevante frequenza nella rete globale.

L'utilizzo di questo approccio permetterebbe di analizzare la rete di interazione genica o di interazione tra proteine, per poi descriverla tramite la composizione di un numero finito di motif, ognuno dei quali andrebbe a ricoprire uno specifico ruolo nella determinazione dell'espressione genica di quella specifica sequenza nucleotidica.

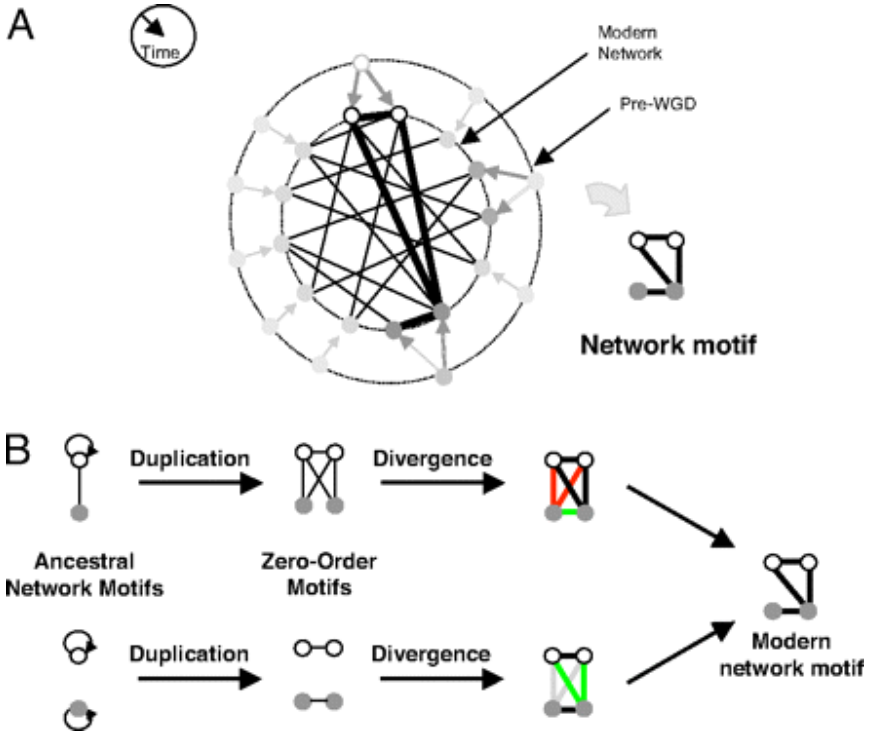


Illustrazione 1: Esempio di motifs riscontrati all'interno della rete di interazione proteica dell'organismo Saccharomyces cerevisiae, dopo il processo di duplicazione.

Come esempio della sua applicazione, si consideri lo studio effettuato sull'analisi della rete di regolazione genica del batterio E. Coli [29] [27],oppure quello relativo all'organismo unicellulare Saccharomyces cerevisiae, il lievito di birra [30].

2.5.2 Caratteristiche del network motif

Considerato il motif come un pattern associato, piuttosto che ad un complesso genico o proteico, alla realizzazione di una specifica funzione cellulare, al cui svolgimento possono concorrere uno o più aggregati molecolari, il suo impiego consentirà, dunque, di stabilire se una particolare entità biologica appartenga o meno ad un insieme di motif già identificato, e di ipotizzare, così, l'attività in cui è coinvolta.

Questi motifs di sottoreti possono presentare, nella rete complessiva, un'estensione variabile e la loro conservazione, tra i vari profili genici dei diversi organismi analizzati, non è detto che si verifichi e nè,tantomeno, in che modo andrà a manifestarsi.

Si analizzino, ora, le due proprietà secondo le quali è possibile definire un network motif[28], ovvero:

- la frequenza, valutata come il numero di volte in cui, nella rete biologica complessiva, sia stato riscontrato il pattern considerato, e di quanto questo valore risulti maggiore rispetto a quello che si era stimato;
- la significatività, in termini statistici, che riguarda proprio quanto sia rilevante la probabilità che un determinato numero compaia, ovvero che questo si verifichi in valore differente a quello stabilito come numero limite.

Generalmente, il valore di significatività statistica di un network motif [31] viene misurato confrontando il numero di riscontri ottenuti alla ricerca della sua occorrenza in una rete biologica, con il numero di ricorrenze rilevate, invece, in una rete casuale.

Per determinare come deve essere misurato il valore di significatività statistica

sono stati effettuati numerosi studi [32] [33], nella maggior parte dei quali vengono utilizzati degli indici statistici a supporto di queste operazioni, quali:

- i "punti zeta" o Z-score, impiegati nel procedimento di standardizzazione di una specifica variabile aleatoria;
- il "valore p" o p-value, che indica la probabilità di ottenere un risultato pari o più elevato di quello effettivamente ottenuto, se è verificata quella che, in statistica, è definita come ipotesi nulla o ipotesi zero, ovvero un'affermazione riguardante la distribuzione di probabilità delle variabili casuali coinvolte nell'analisi.

Immagine 19: Esempio di tutti i possibili modelli di interazione che possono instaurarsi tra due proteine.

Tuttavia, nonostante gli approcci più comunemente adottati applichino il concetto di motif contestualizzandolo, soltanto, in relazione alla topologia che presenta la rete biologica, questa sua caratterizzazione risulta inadeguata in alcuni ambiti e questo ha condotto all'elaborazione di una nuova definizione di "motivo di rete", che tenga conto, oltre che della struttura della rete, del ruolo funzionale che svolgono i componenti identificati come appartenenti ad un motif, valutandolo in un contesto di applicazione che riguarda un tipo specifico di rete biologica, le reti metaboliche [34].

2.5.3 Tecniche basate sul network motif

Tra gli approcci orientati all'identificazione di questi sottografi ricorrenti, basati sulla topologia che caratterizza la rete biologica su cui effettuare la ricerca, è

da riportare, necessariamente, quello utilizzato da Shenn-Orr per l'analisi della rete di regolazione genica del batterio E. Coli [27].

Nella rete analizzata, è stato possibile individuare la presenza di tre differenti motivi di rete, costituiti ciascuno da tre oppure quattro proteine, assume particolare rilevanza il "Feed Forward Loop" (FFL [35]), un motif caratterizzato dall'aver tre interazioni di trascrizione per questo categorizzato come "Three-Protein Network Motif", la cui presenza è stata, in seguito, riscontrata anche nella rete di regolazione genica del lievito di birra.

Nello specifico, questo motif è costituito da un fattore di trascrizione X, che regola un secondo fattore di trascrizione Y, ed entrambi si occupano di definire la regione di regolazione che interessa il gene Z, ovvero il fattore target, calibrando congiuntamente la velocità di trascrizione.

La tecnica finora descritta è poi stata rispresa ed estesa da altre euristiche, come, ad esempio, l'approccio che considera i motifs come distinguibili, fondamentalmente, sulla base di due tipi di interazioni [36], ovvero le relazioni tra proteine e quelle che riguardano la correlazione che sussiste tra la regolazione e la trascrizione genica, identificate, nelle reti, tramite gli archi contrassegnati da un colore differente.

Per quanto riguarda, invece, i metodi che sono stati elaborati tenendo conto che il contesto di applicazione consiste nell'occuparsi della ricerca dei network motifs proprio nelle reti biologiche, si possono citare, a titolo esemplificativo:

- l'algoritmo denominato Qcut che, per identificare i complessi proteici, esegue un raggruppamento incentrato sulla valutazione della modularità che questi dimostrano, ovvero assegnando ad uno stesso cluster tutti gli aggregati che risultano essere inferiori, in numero, ad una determinata soglia, oppure che si vedranno essere caratterizzati da un'elevata densità tra gli elementi che li compongono. I clusters che sono stati

creati verranno poi, ricorsivamente, scomposti in dei sottogruppi, in modo estrarre una sub-community di interesse [37];

- l'algoritmo di clustering Markov (MCL), che applica una tecnica di clustering non esclusiva, ovvero più flessibile in termini di classificazione dei dati, poichè permette che un elemento della rete possa appartenere, contemporaneamente, a più clusters, se caratterizzato da un diverso grado di appartenenza. Questo tipo di approccio è denominato anche fuzzy clustering e nel caso dell'algoritmo MCL, viene implementato utilizzando il concetto di cammino casuale o "random walk" su grafo, proprio per ricercare gli elementi che andranno a costituire i clusters [38];
- gli algoritmi genetici o "Genetic Algorithms", impiegati su reti di interazione tra proteine, per identificare dei clusters distinti in base alle funzioni cellulari svolte del complesso proteico che vi appartiene. Secondo questo tipo di approccio, il grafo che rappresenta la Protein-Protein Interaction Network, deve essere rappresentato attraverso una matrice di adiacenza, che descriva, come set di dati, i campioni genici di un determinato individuo, i quali andranno a costituire proprio i nodi della suddetta rete. Relativamente al criterio adottato per effettuare la clusterizzazione si avrà che, considerato un generico gene i ed un qualsiasi altro gene g e che esista un arco nel grafo che rappresenti un legame funzionale tra i due, ove il suddetto arco, nella matrice di adiacenza, sarà rappresentato da un valore j, entrambi risulteranno essere identificati come appartenenti allo stesso cluster [39] .

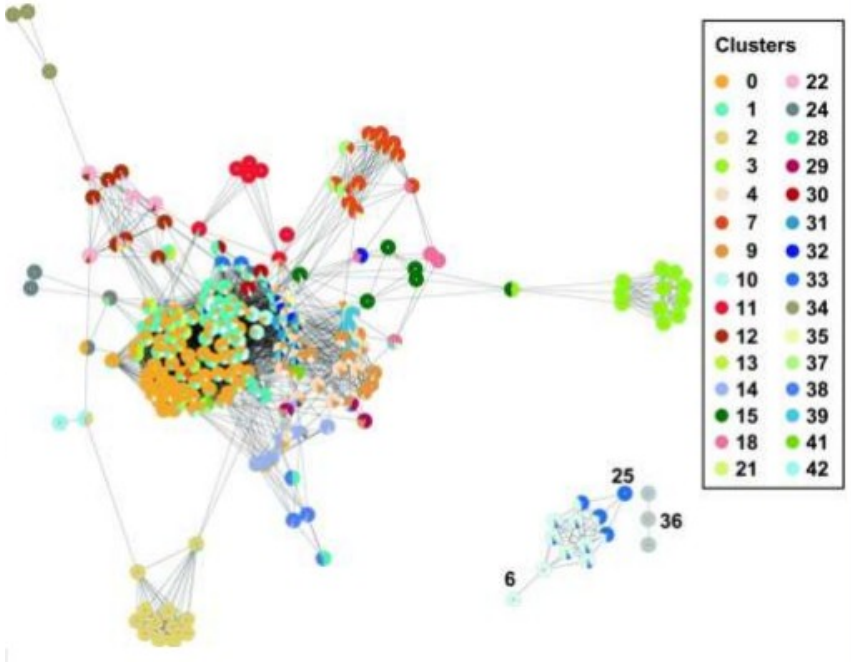


Immagine 20: Identificazione dei clusters di una rete genica di un batteriòfago ottenuta con l'algoritmo di clustering Markov.

Capitolo 3: Pattern Discovery

3.1 Data Mining

Lo studio di queste reti biologiche comporta, necessariamente, il dover fronteggiare la gestione di un'elevata mole di dati, la cui elaborazione ha, come obiettivo, quello di estrarre informazioni di interesse, a prescindere dalla natura che presentano le risorse di cui si dispone o dalla quantità in cui queste vengono presentate.

Tuttavia, il problema relativo alla gestione di ingenti quantità di dati, riguarda qualsiasi ambito culturale e non soltanto quello correlato all'esaminare i geni che costituiscono le sequenze nucleotidiche degli organismi, perciò la definizione di tecniche e metodologie atte alla sua risoluzione è identificata dall'appellativo di Data Mining.

3.1.1 Definizione di Data Mining

Per quanto riguarda l'ambito informatico, con il termine "Data Mining" si intende l'analisi di un database di dimensioni considerevoli, attraverso l'ausilio di metodi matematici, che non si limiti soltanto all'identificazione ed alla classificazione dei dati, ma che sia orientato anche verso lo sviluppo di metodi automatizzati che si occupino del filtraggio, della selezione e dell'interpretazione di questo insieme eterogeneo di nozioni, tramite un processo di elaborazione che permetta di stilare delle statistiche e di riscontrare se esistano o meno delle correlazioni tra le informazioni che sono state estratte.

È opportuno riportare, allora, la differenza che sussiste da il concetto di dato e

quello di infromazione, ovvero:

- si definisce "dato" un generico elemento, espresso in forma simbolica, le cui caratteristiche risultano essere note, ma non ancora rielaborate e quindi catalogate;
- si definisce "informazione" il risultato ottenuto dal processo di analisi e classificazione di un determinato dato, ovvero l'estrazione del potenziale informativo che è racchiuso nel suo contenuto, prima ancora di essere elaborato;

L'attuazione di questa trasfomazione della risorsa grezza, rappresentata dal dato, in prodotto finito, quale l'informazione che si è ricavata dalla sua analisi, comporta l'acquisizione di conoscenza, ovvero la raccolta di una serie di informazioni che, aggregate tra loro, permettano di arricchire il patrimonio culturale, in termini di esperienza e comprensione.

Inizialmente, la disciplina che si occupava di perseguire e realizzare questo obiettivo era la statistica, ma con lo sviluppo e la diffusione delle tecnologie dell'informazione si è manifestata l'esigenza di ricercare ed archiviare un'ingente quantità di informazioni di varia natura, imbattendosi, di conseguenza, nel cosiddetto "problema dell'esplosione dei dati", tanto che John Naisbitt scrisse, nel suo libro intitolato "Megatrends: Ten New Directions Transforming Our Lives", pubblicato per la prima volta nel 1982, la seguente frase: "Stiamo annegando in un mare di informazioni, tuttavia siamo affamati di conoscenza".

A questo punto, considerando la velocità di crescita che caratterizza il continuo aumento della quantità di informazioni memorizzata sui supporti di archiviazione, l'inadeguatezza che dimostrano le tecniche di elaborazione tradizionali e la potenza di calcolo raggiunta dai sistemi con lo sviluppo delle tecnologie informatiche, si è maninfestata l'esigenza di una disciplina che

soddisfacesse questa ricerca di conoscenza, ovvero il Data Mining, che rappresenta proprio il fulcro del processo di Knowledge Discovery nei Databases (KDD).

3.1.2 Tecniche di Data Mining

Esistono svariate tecniche che si basano sull'applicazione del concetto di Data Mining, che differiscono in base al fine per cui sono state progettate e di conseguenza, anche dal punto di vista implementativo, tuttavia, risultano essere tutte strutturate seguendo una procedura standar che descrive il modus operandi da seguire, affinchè i dati di cui dispone vengano adeguatamente analizzati.

Dunque, un generico processo di data mining [40] è descritto dal seguente insieme di azioni:

1. Definizione degli obiettivi che si vogliono ottenere dall'analisi;
2. Ricerca e recupero delle informazioni di interesse;
3. Preparazione dei dati, tramite delle azioni di :
    - Pulizia, per ridurre il rumore ed eliminare degli eventuali errori,
    - Arricchimento, integrando le risorse di cui si dispone con delle ulteriori sorgenti di informazione;
    - Codifica, secondo un determinato criterio, in base al modello in cui si vuole che i dati da analizzare vengano descritti;
4. Applicazione dell'algoritmo di data mining, che si occupa proprio dell'analisi dei dati e della conseguente estrazione delle informazioni;
5. Preparazione e restituzione dei risultati;

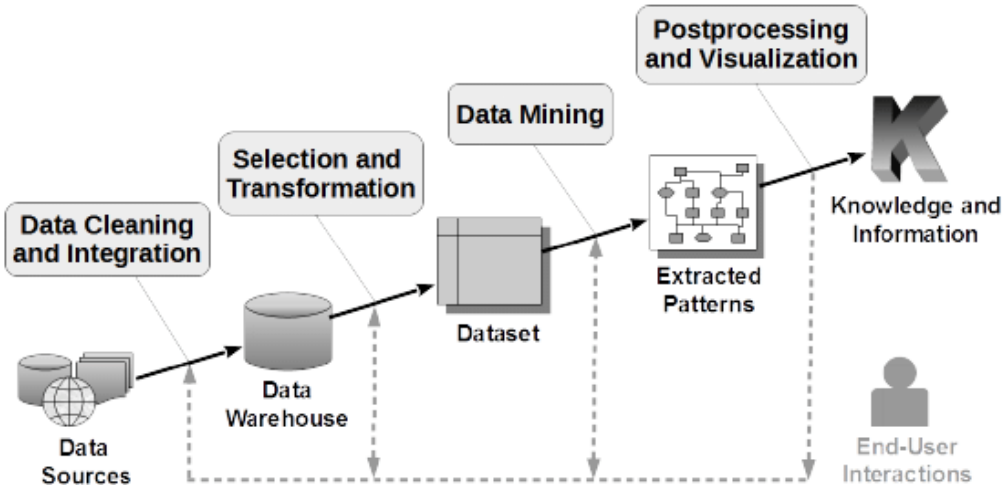


Immagine 21: Fasi principali di un processo di data mining.

Per quanto concerne all'implementazione dell'algoritmo che mette in atto l'effettiva fase di data mining, questa varia in base alla tecnica che si è deciso di adottare ed in relazione al tipo di dati che devono essere processati.

Tali tecniche si suddividono nei seguenti tipi:

- di regressione, che utilizzano la conoscenza acquisita per classificare nuovi elementi;
- di clustering, che permettono di individuare dei raggruppamenti di dati, che manifestano delle regolarità comuni e di distinguerli, così, da altri gruppi;
- basate sull'identificazione di associazioni o di sequenze ripetute, per determinare il motivo per cui si verifichino queste occorrenze;
- di classificazione;

Qualunque sia l'approccio che si sceglie di adottare, il processo di estrazione della conoscenza avverrà tramite la ricerca e l'identificazione di

corrispondenze, riscontrate nei dati da analizzare, a cui ci si riferisce con il nome di "pattern", ovvero un modello di cui si verificano delle occorrenze.

3.2 Pattern Mining

Per "pattern mining" si intende, allora, la ricerca di una specifica struttura all'interno di un insieme di dati, in modo da verificare se tra questi si manifestano delle regolarità che è possibile descrivere ed identificare attraverso la definizione di un modello.

3.2.1 Caratteristiche del patter mining

Se si considera il pattern mining per quanto concerne l'analisi di dati formalizzati tramite dei grafi, allora l'obiettivo è quello di identificare delle sottoreti isomorfiche, che risultino essere significativamente ricorrenti e che presentino sempre le stesse caratteristiche.

Comunque, anche riferendosi ad un generico contesto applicativo di pattern mining, queste vengono valutate seguendo alcune metriche definite proprio per verificare il soddisfacimento delle regole associative che possono manifestarsi in un set di dati.

Dunque, gli aspetti di cui si deve tener conto, quando si effettua un'analisi di questo tipo, sono i seguenti:

- la transazione che si sta analizzando, per determinare se si verifica una situazione per cui la presenza di un elemento nel set di interesse risulta essere strettamente dipendente da quella di altri elementi, appartenenti, anch'essi, allo stesso insieme di dati. In questo caso, allora, la suddetta

transazione rappresenta, di per sè, costituita da un sottoinsieme di informazioni correlate tra loro;

- la frequenza con cui ricorre il pattern, in un determinato numero di transizioni, calcolando la percentuale con cui si verificano queste occorrenze e verificando se il suo valore sia superiore ad una soglia fissata, detta "supporto";
- il supporto o "support" di un modello m, definito come il numero di transazioni in cui si è potuta verificare la presenza di questa struttura;
- la confidenza o valore di "confidence", che definisce quante volte il pattern ricorra nelle transazioni di cui si è già stabilito che lo contengano;

Questo approccio di ricercare dei patterns ricorrenti discende dal "Frequent Item Set Mining", un metodo utilizzato nella "Market Basket Analysis", che si occupa di analizzare le abitudini di acquisto adottate dai clienti nella vendita al dettaglio, al fine di determinare delle associazioni tra i differenti prodotti che vengono comprati.

Immagine 22: Esempio di identificazione delle associazioni che sussistono tra i prodotti acquistati insieme, poichè, secondo la Market Basket Analysis, il 98% dei consumatori che si troverà a comprare i prodotti A e B, alla fine acquisterà anche C.

3.2.2 Algoritmi di pattern mining

L'area di ricerca in cui agisce il pattern mining è piuttosto vasta, perciò si effettua una distinzione tra le differenti tecniche che sono state elaborate, in base all'obiettivo che si prefiggono di raggiungere.

È possibile definire, allora, quattro diverse categorie, dedite a:

- l'elaborazione di algoritmi sempre più efficienti, che si occupino di estrarre delle sequenze ricorrenti, implementati secondo molteplici strategie, sia per quanto riguarda la consultazione e la rappresentazione dei dati, sia per quanto concerne l'effettiva generazione di contenuto;
- la risoluzione dei problemi riscontrati per quanto riguarda la scalabilità degli stessi dati;
- il proporre varianti di algoritmi di estrazione di conoscenza, capaci di agire su domini di dati differenti, dal punto di vista di come è attuata la loro rappresentazione, e che permettano di ricercare, su questi, modelli che varino in base all'informazione di interesse;
- il concretizzare dei metodi applicativi che consentano di utilizzare, agevolmente, tutte le tecniche elaborate per il data mining [41].

Proprio in questo scenario si collocano, quindi, tutte le euristiche che discendono da quello che è il voler determinare, in maniera efficiente, la presenza di elementi ricorrenti, nelle transazioni che si verificano in un insieme di dati assegnato.

A queste ci si riferisce con il nome di frequent pattern mining ed includono metodologie come:

- il sequential pattern mining, che, ad esempio, identifica come "modello sequenziale" il verificarsi, nella transazione di un database, dell'acquisto

di una fotocamera, seguito da quello di un dispositivo di archiviazione di massa, come una memory card.

- lo structural pattern mining, che riconosce, come "struttura ricorrente", sia questa rappresentata come un sottografo o una sottosequenza, un raggruppamento di elementi che possano essere accomunati ed identificati come un unico insieme di dati, detto "itemset";
- il correlation mining;
- la classificazione associativa;
- il clustering basato sui patterns ricorrenti.

Questo approccio fu proprosto da Agrawal, nel 1993, proprio nell'ambito che riguarda le transazioni che si verificano in un database ed è stato esteso ed implementato in numerose varianti, alcune delle quali verranno descritte qui di seguito.

3.2.3 Frequent Pattern Mining

Si definisce "Frequent Pattern Mining" il processo di estrazione di conoscenza che avviene andando a ricercare i modelli o le strutture che ricorrono più frequentemente nei database che costituiscono le risorse da elaborare.

Questi modelli possono essere dei singoli elementi oppure dei raggruppamenti o sottostrutture costituite da questi ultimi, che appaiono, nel set di dati considerato, con una frequenza che si dimostra essere pari o superiore ad un valore di soglia, stabilito prima di effettuare l'analisi [42].

Possono essere identificate tre metodologie di base che è possibile adottare nell'estrazione di itemsets ricorrenti e nella conseguente definizione delle

regole associative che descrivono questa proprietà, quali :

- l'algoritmo "A-priori";
- l'algoritmo FP-growth;
- l'algoritmo Eclat.

Ognuna di queste verrà analizzata, soffermandosi sul meccanismo su cui si fonda ed il relativo contesto applicativo.

Algoritmo Apriori

Questo algoritmo costituisce un metodo di ricerca delle associazioni, utilizzato nel data mining, per generare degli itemsets frequenti, ovvero dei raggruppamenti ricorrenti, partendo dall'analisi di itemsets che sono, invece, composti da un elemento soltanto.

La proprietà su cui si fonda questo approccio, riconosciuta da Agrawal e Srikant nel 1994, è la seguente:

"Considerato un insieme di oggetti, la cui ricorrenza è stabilito essere frequente, allora tutti i sottogruppi che è possibile identificare nel sudddetto dataset saranno, a loro volta, ricorrenti".

Questo presupposto, noto come "principio di anti-monoticità" comporta, dunque, che un itemset identificato come non ricorrente, sarà composto da dei sottogruppi che, di conseguenza, non saranno frequenti.

L'algoritmo in questione, allora, per ricavare le associazioni, costruisce gli itemsets frequenti aggiungendovi un elemento per volta, dopo aver stabilito quale, tra i candidati che sono stati generati, risulta essere idoneo per essere contrassegnato come frequente. Un riscontro a quanto stabilito verrà poi ricercato effettuando una comprova sul dataset e le iterazioni che l'algoritmo compie terminano soltanto quando non si riesce più ad estrarre dall'insieme

dei dati, degli elementi di cui può verificare la ricorrenza.

Al fine di ridurre il numero di iterazioni eseguite, dovute alle scansioni che devono essere effettuare ai dati per ricercare i patterns frequenti, sono state elaborate diverse versioni di questo algoritmo, utilizzando vari approcci implementativi, come l'utilizzo di tecniche di hashing [43], di partizionamento [44] o di campionamento [45] .

Algoritmo FP-growth

Questo algoritmo è stato ideato per tentare di risolvere i problemi in cui incorre la strategia A-Priori, ovvero il dover affrontare costi computazionali elevati a causa della generazione di ingente numero di candidati, tra i quali estrarre, ogni volta, quello da inserire nell'itemset, nonostante questi vengano ridotti notevolmente rispetto a quanto effettuato da una qualsiasi tecnica di tipo greedy e della necessità di dover esaminare, ad ogni iterazione, l'intero database.

La metodologia che caratterizza l'FP-growth[46], invece, è radicalmente diversa, poichè la determinazione degli itemset frequenti avviene senza che sia generato alcun candidato e di conseguenza, senza che sia necessario effettuare alcuna verifica che accerti che il candidato estratto sia, effettivamente, un frequent item.

Questo è possibile tramite l'impiego di una struttura dati chiamata "albero dei pattern frequenti" o "FP-tree", che conserva, di volta in volta, le informazioni associative sull'itemset che si sta costruendo.

La particolarità risiede nel fatto che, già con la prima scansione del database, vengono selezionati soltanto gli elementi identificati come frequenti, perchè caratterizzati da un valore di supporto pari o superiore alla soglia prestabilita, escludendo, dunque, tutti gli altri.

Le dimensioni dell'FP-tree verranno, allora, ridotte notevolmente, pur

conservando, tuttavia, tutte le informazioni di rilievo atte a poter riconoscere ed estrarre, agevolmente, i patterns frequenti.

Inoltre, l'algoritmo è strutturato in maniera da identificare degli itemset più lunghi, sfruttando la ricerca, attua ricorsivamente, di patterns frequenti che sono, invece, di dimensioni ridotte rispetto ai precedenti, che verranno poi aggiunti, nell'albero FP, come concatenazione di questi.

Si può concludere, quindi, che rispetto a quanto ottenuto dall'implementazione della logica Apriori, il tempo impiegato dall'algoritmo FP-growth nella ricerca viene ridotto in maniera considerevole.

Tra le alternative che sono state proposte, relativamente a questo approccio, si cita quella di Agarwal, del 2001, che si focalizza sulla ricerca di itemsets frequenti utilizzando una visità dell'albero in profondità e quella di Grahne e Zhu, del 2003, basata su un'implementazione tramite array.

Eclat

L'algoritmo di equivalenza CLASS Transformation, denominato anche Eclat, proposto da Zaki nel 2000, si distingue nettamente dai due approcci descritti in precedenza, nonostante rappresenti, di fatti, una variante della metodologia Apriori, poichè permette di ispezionare i dati in una maniera del tutto differente.

Sia l'Apriori che L'FP-growth, infatti, estraggono gli itemsets frequenti da un database organizzato secondo un modello orizzontale, mentre Eclat, in alternativa, memorizza la lista delle transazioni su cui deve lavorare in maniera verticale, così come mostra l'immagine sottostante [47].

Transazione	Beni
transaz1	latte, uova, birra
transaz2	pane, pasta
transaz3	yogurt, pane, pizza, latte
transaz4	uova, pane

Organizzazione orizzontale (Apriori)

Immagine 23: Organizzazione dei dati secondo un modello orizzontale, come attuato per l'algoritmo Apriori.

latte	uova	birra	pane	pasta	yogurt
200	100	200	300	400	100
150	200	100	200	100	150
		100	50		

Immagine 24: Organizzazione dei dati in maniera verticale, utilizzato dall'algoritmo Eclat.

La classe tidList viene utilizzata dalla classe itemets per memorizzare gli ID della transazione che supportano ciascun elemento minato. Solo l'implementazione dell'algoritmo di estrazione Eclat produce ID transazioni. tidList utilizza la classe itemMatrix per memorizzare in modo efficiente gli elenchi ID transazioni come una matrice sparsa.

La prima scansione del database si occupa di costruire la TID_set di ogni singolo elemento, ovvero una struttura che memorizzi gli ID associati alle transazioni che interessano il suddetto.

Gli itemsets frequenti verranno poi generati, una volta che è stato identificato il primo frequent pattern, applicando la proprietà a-priori, senza però avere la necessità di iterare nuovamente il database per verificare la correttezza dell'operazione, controllando il valore di support dell'elemento riconosciuto come frequente, perchè nel TID_set sono incluse anche informazioni di questo tipo. Per quanto riguarda l'ordine con cui vengono scansionati gli elementi, questo segue un metodo di visita in profondità, simile all'approccio FP-growth ed inoltre, l'algoritmo termina fino a quando si vagliano tutti i possibili frequent patterns oppure quando non sono più presenti candidati da valutare nella lista che si è generata.

Un altro lavoro correlato a questo tipo di tecnica è quello intrapreso da Holsheimer, nel 1995, che ha dimostrato, in generale, come la scelta adeguata della struttura dati possa incidere in maniera positiva sulle performance di questo tipo di algoritmi.

3.2.4 Frequent Pattern Mining su reti biologiche

Il frequent pattern mining può essere applicato anche nell'ambito delle reti biologiche, puntando all'eleborazione di tecniche che si focalizzino sull'identificare la ricorrenza di patterns, rappresentati, in questo caso, da sottografi, caratterizzati da un insieme di nodi e di relazioni tra essi, che si ripresentano, più e più volte, nella rete che descrive l'insieme di campioni che si vuole analizzare.

L'intento di voler ricercare questi patters, definiti come "emergenti"[48], ognuno dei quali costituisce una struttura che si presenta con frequenza elevata in una specifica regione della rete, è volto ad evidenziare le differenze che sussistono tra un insieme di individui ed un altro, identificando, così, le peculiarità che permettano di comprendere le relazioni che descrivono le attività cellulari e come queste si manifestino nell'espressione genica.

L'utilizzo degli "emerging patterns" si basa, in genere, su algoritmi che impiegano tecniche matematiche, come un metodo di discretizzazione che tiene conto dell'entropia [49],in modo da etichettare i geni che risultano essere coinvolti nell'insorgere di una specifica malattia, a causa della loro manifestazione fenotipica nell'individuo affetto.

Un altro approccio è quello di ricercare dei "discriminative patterns" ovvero dei sottografi speciali, discriminanti, poichè riscontrati come significativi, al fine di comprendere le varie differenze riscontrabilitramite il confronto dei profili genici di pazienti malati, con le sequenze nucleotidiche di soggetti che, invece, non

sono affetti dalla patologia di interesse.

Nella valutazione, si tiene conto di alcune regole statistiche, come la valutazione del valore di confidenza o "confidence", che descrive la probabilità di riscontrare la presenza del pattern in esame, quando è già stato individuato un modello ad esso correlato.

Il perseguire questo metodo potrebbe essere utile in un contesto applicativo in cui si incorra nella necessità di dover classificare gli elementi analizzati, siano essi aggregati molecolari oppure interi profili genici, in gruppi differenti, ai fini di raccogliere i dati in categorie qualitativamente contrastanti, evidenziando, così, le differenze che li caratterizzano.

Per analizzare l'origine di una determinata malattia, è comune, ad esempio, classificare i pazienti, dei quali si hanno a disposizione i campioni relativi alle loro sequenze nucleotidiche, in due raggruppamenti differenti [50], l'insieme degli individui malati e l'insieme degli individui, che invece, non manifestano la patologia.

Un'altra metodologia, atta a semplificare la complessità computazionale, consiste nell'adottare una tecnica ad-hoc, che si prefigga di utilizzare il profilo genico di un individuo che manifesta delle singolarità, ovvero le eccezionalità che si vogliono identificare nel database da analizzare, proprio come target da andare a ricercare nella rete biologica, in modo da estrarre, dal resto della popolazione esaminata, altri soggetti che siano accumunati dalle stesse peculiarità, per quanto riguarda l'espressione del loro corredo genetico.

In ogni caso, il criterio adottato dalla pattern discovery, ovvero la tecnica che si occupa della ricerca di un modello, sia questo speciale, ricorrente o emergente, all'interno di una determinata rete da analizzare, dipende strettamente dall'esigenza che vuole soddisfare il committente dell'analisi, e quindi anche da cosa si intende, in dipendenza dall'ambito applicativo, per

pattern significativo.

3.2.5 Estrazione di informazioni biologiche tramite la pattern discovery

In generale, la volontà che accumuna tutte le tecniche di pattern discovery su reti biologiche che sono state eleborate finora è quella di cercare di identificare e comprendere quali complessi proteici svolgano un ruolo fondamentale nello svolgimento delle attività cellulari che interessano gli organismi viventi, cercando di individuare quali di questi si siano mantenuti invariati nel corso del processo evolutivo e quali funzioni possano essere attribuite alle entità il cui compito risulta essere ancora sconosciuto, in correlazione, sempre, alle similarità che queste dimostrino al confronto con altri aggregati genici.

I dati da analizzare, però, devono essere formalizzati adottando una rappresentazione tramite grafi, in modo che si riescano a descrivere differenti tipi di informazioni biologiche, con completezza.

Si devono, infatti, costruire delle reti che raffigurino, tramite degli archi, le associazioni di interazione che si creano tra le diverse entità biologiche che costituiscono una cellula, siano queste relative ad interconnessioni fisiche per l'attuazione di attività biochimiche oppure al fenomeno di coespressione genica.

Infatti, la direzione verso cui sta procedendo la ricerca scientifica, vista la vastità di dati che si è riusciti a raccogliere sulle seguenze geniche degli acidi nucleici, è proprio l'acquisizione di quanta più conoscenza possibile nell'ambito medico, per tentare di elaborare un strategia che consenta di accertarsi di quali siano le entità molecolari che possano contribuire o, in qualche modo, influenzare il presentarsi di una specifica disfunzione, il decorso di una malattia, o gli effetti che conseguono ad una particolare mutazione genica.

Come dimostrato da Barabasi, Gulbahce e Loscalzo, è fondamentale,

nell'analisi di una qualsiasi rete biologica e specialmente per quanto riguarda le reti di interazione geniche e tra proteine, focalizzarsi sugli effetti che manifestano, nel corredo genetico di un qualunque individuo, a causa delle correlazioni che si intestaurano tra il suo genotipo ed il suo fenotipo [51].

Questo studio ha evidenziato, allora, l'inadeguatezza che comporta la valutazione di una rete biologica basandosi esclusivamente sulla topologia che la caratterizza.

È da considerarsi una prova di quanto affermato il lavoro portato avanti per quanto concerne l'indagine volta a determinare le relazioni che intercorrono, quindi, tra i geni e le malattie, poichè è risultato che la maggior parte dei geni che, effettivamente, siano da considerarsi come responsabili dell'insorgere di queste, non sarebbero potuti emergere tramite un'analisi della topologia della rete, poichè in questa non appaiono in posizioni centrali e quindi correlate a funzioni di rilievo ma, al contrario, sono localizzati in regioni piuttosto periferiche e marginali [52].

Questo perchè, nonostante l'insorgere delle malattie sia causato da mutazioni geniche o disfunzioni cellulari, se queste risultano essere non letali per l'organismo che ne è affetto, allora non è possibile che abbiano intaccato funzionalità che, invece, sono indispensabili per la sua sopravvivenza.

Ciò ha condotto alla formulazione di un'ipotesi riguardo la localizzazione dei geni che causerebbero la crescita eccessiva e scoordinata delle masse tumorali, basata sul fatto che essendo questi coinvolti proprio nelle attività principali che concernono la vita di una cellula, ovvero la sua crescita e duplicazione, allora è probabile che si trovino in quegli aggregati molecolari che manifestano una certa rilevanza e centralità, in termini di rilevanza funzionale e topologica [53].

Per quanto riguarda il software utilizzato nella parte di sperimentazione

riportata in questo elaborato, l'algoritmo implementato è indirizzato all'indentificazione di patterns discriminanti e quindi di sottografi ritenuti significativi poichè rilevanti per effettuare un discernimento dei dati e definire le due categorie da mettere a confronto.

Considerate queste due classi di informazioni, infatti, verranno generati due grafi che rappresenteranno, rispettivamente, il set di geni presenti nei campioni dei pazienti malati e quello estratto dai campioni appartenenti ai soggetti sani, per poi tracciare delle corrispondenze che esplichino le correlazioni esistenti, in termini similarità funzionali e topologiche, che si riscontrano tra le due reti.

Questa metodologia di "discriminative pattern mining" è ampiamente impiegata nella bioinformatica, perciò è stata estesa e rivisitata in base alla tipologia di dati su cui doveva essere applicata ed al tipo di risultati che si voleva fossero ottenuti.

Ad esempio, l'estrazione dei sottografi discriminanti può essere attuata effettuando dei tagli in maniera tale da prelevare soltanto quelli che soddisfano dei criteri stabiliti [54] prima di eseguire l'analisi, come il dover rispettare un valore minimo di supporto, oppure, come accade nella tecnica di ricerca dei "synergy graph patterns" [55], calcolando una sorta di punteggio che descriva il potere discriminante della struttura in esame, utilizzando il valore di confidence.

L'approccio che più si avvicina a quello utilizzato dall'algoritmo del software, però, è quello adottato nella procedura di analisi atta a misurare i processi biologici di alcune cellule per evidenziare quali geni siano coinvolti o meno, dal punto di vista funzionale, nello sviluppo di neoplasie [56].

Lo scopo consiste nel costruire un database in cui ricercare, dopo aver raggruppato i dati in base alla presenza o meno di attività mostrata nel processo cellulare di interesse, dei "dissimilar graph patterns", costruendo delle relazioni tra i due insiemi, in base ad un rapporto di correlazione, valutato

in termini di similarità dal punto di vista strutturale e di significatività statistica.

Capitolo 4: Discriminating Graph Pattern Mining

Il genoma umano è organizzato in maniera tale da permettere il corretto svolgimento di molteplici funzionalità, perciò il poter analizzare la sua struttura, tramite l'ausilio di svariate tecniche,come quelle di nuova generazione e dei microarrays, ha permesso di estrapolare ingenti quantità di dati da cui poter acquisire informazioni.

Ad esempio, è stato possibile verificare, scansionando differenti corredi genomici, la presenza di raggruppamenti o clusters di geni manifestanti una similare tipologia di espressione e localizzati in regioni cellulari contigue o addirittura nello stesso complesso proteico, poichè accumunati dall'aver attraversato la stessa storia evolutiva [57].

Allora, l'esigenza di comprendere il significato che si cela dietro l'instaurarsi di questi legami genici e la funzionalità che questa correlazione possa permettere di realizzare all'interno di un essere vivente, ampliare e condividere, quanto più possibile, i database genomici che sono stati creati proprio per questo scopo.

4.1 Introduzione all'algoritmo utilizzato

Il software utilizzato si propone di risolvere il problema dell'estrazione di patterns eccezionali in reti di tipo biologico, che si occupano di raffigurare un

set di dati relativi all'espressione ed all'interazione genica, sulla base delle sequenze nucleotidiche estratte da alcuni campioni [10].

Lo scopo, come si è già preannunciato, è quello di identificare delle singolarità tra gli aspetti che caratterizzano i profili genici dei dati ricavati tramite l'analisi di individui malati, in modo da differenziarli da quelli che risultano essere propri, invece, del corredo genetico dei soggetti sani.

I parametri che verranno valutati per stabilire le relazioni che si instaurano tra due o più geni riguardano le similarità locali, cioè se questi si presentano nelle sequenze nucleotidiche mantenendo lo stesso ordine nella disposizione che assumono, tenendo conto anche degli effetti che ne conseguono.

Questo tipo di approccio richiede che i dati vengano rappresentati secondo un modello specifico, in modo tale da poter effettuare il processo di mining attuando una tecnica improntata alla ricerca dei patterns sui grafi.

Allora, la rete che modellerà i dati sarà un grafo non orientato e connesso, i cui nodi, etichettati tramite degli id, andranno ad identificare i geni presenti nel profilo genico o nell'attività biochimica che si vuole analizzare, mentre le relazioni tra questi verranno esplicate tramite degli archi pesati.

Sulla base di questo modello di rete, il potere discriminante di un pattern o sottografo eccezionale verrà misurato prendendo in esame la rilevanza che caratterizza la coespressione tra ogni coppia di geni che fa parte del grafo.

4.2 Struttura della rete

I dati da formalizzare riguardano le informazioni estratte da campioni di individui, relativamente all'espressione genica, ovvero la misura del livello in cui vari geni si manifestano nelle sequenze nucletodiche di un determinato dataset.

Il profilo genomico di ogni paziente, allora, andrà ad essere rappresentato tramite un grafo G = (V,E), che sia etichettato, non orientato e connesso, poichè per ogni coppia di nodi W e Y dovrà esistere almeno un cammino che colleghi il vertice W ad Y.

Un generico vertice rappresenta un gene per cui definire le relazioni in cui si trova ad essere coinvolto ed il livello di espressione con cui si manifesta rispetto ad altri, tramite l'insistenza di un arco tra tutti i complessi proteici coinvolti nella specifica interazione.

Ai fini di estrarre i patterns discriminanti, è necessario che si disponga di attributi che vadano a qualificare i dettagli e le caratteristiche di queste correlazioni geniche, perciò, in questa struttura, gli archi saranno pesati.

Relativamene ad ogni coppia di geni che risulta essere conessa tramite un arco, dunque, verranno espressi due tipi di pesi, definiti in termini di "robustezza" e "rilevanza", per esplicitare la natura della relazione che sussiste tra i due.

4.2.1 Strength

Per "strength" si intende la robustezza che caratterizza la relazione che si instaura tra due geni, propri del profilo genetico di un individuo t, che appartiene alla popolazione di campioni dalla quale sono stati estratti i dati.

Questa correlazione, che si riferisce all'interazione che sussite tra la coppia di geni identificati come a_i e a_j, rappresenta la massimizzazione del valore di probabilità di riscontrare, tramite osservazioni empiriche, la loro coespressione genica. Questo significa che nel paziente t, al quale appartengono le sequenze

nucleotidiche in esame, si manifesteranno, contemporaneamente, sia a_i che a_j.

4.2.2 Relevance

Con il termine "relevance" ci si riferisce alla misurazione della probabilità che si ha di riscontrare un elevato valore di robustezza durante l'analisi della correlazione che sussiste tra la coppia di geni a_i e a_j considerata.

L'intento di voler utilizzare, come peso dell'arco che esplica la relazione genica, questa "rilevanza" è motivata dal fatto che il livello di espressione di uno specifico gene vada a condizionare il valore della correlazione che andrà a caratterizzare una possibile interazione.

Ne consegue, quindi, che più il livello di espressione di un gene è alto, più aumenta la probabilità che tra questo ed un altro gene si instauri un rapporto di correlazione significativo, in termini di rilevanza, che non sia invece frutto della casualità.

4.2.3 Costruzione del modello

Sulla base dei parametri sopradescritti si è scelto di costruire un modello di rete che li utilizzi, dunque, come pesi da associare a ciascun arco tracciato tra le varie coppie di nodi.

Ogni "Strength-Relevance Network" sarà un grafo che, adottando la struttura proposta, si occupi di descrivere il profilo genico associato ad un determinato individuo t facente parte della popolazione di campioni da sottoporre all'agoritmo.

Per rendere i dati idonei alla tecnica di ricerca di patterns discriminanti

implementata dal software, si deve definire un grafo per ogni individuo t di cui esaminare il corredo genetico, distinguendo nel database della popolazione a cui appartengono i campioni, quali di questi siano sani e quali, invece, malati, in modo tale da descrivere una rete che sia unica per ogni paziente e che tenga traccia del suo stato medico.

Inoltre, per quanto riguarda le informazioni memorizzate come pesi degli archi della SR-Network, queste andranno ad essere confrontate con dei termini di riferimento, ovvero dei parametri di soglia in base ai quali effettuare dei calcoli computazionali per determinare se tra quella specifica coppia di geni esista un'interazione da considerare.

Questi valori di riferimento saranno indicati, rispettivamente, come :

- il valore di soglia τ_s , per quanto riguarda il valore di "strength threshold", ovvero la robustezza che caratterizza il livello di coespressione tra una coppia di geni a_i e a_j;
- il valore di soglia τ_r , riferendosi al valore di "relevance threshold".

Allora, sarà possibile inserire un arco tra i due geni a_i e a_j se la relazione che sussiste tra questi è caratterizzata da una robustezza che si presenta in valore superiore alla soglia τ_s ed un valore di rilevanza anch'esso maggiore di quello fissato per il parametro τ_r .

4.3 Definizione ed approccio al problema

Siccome l'obiettivo che si vuole raggiungere analizzando i campioni di cui si

dispone è quello di ricercare ed identificare, se presenti, dei patterns che rappresentino delle peculiarità ricorrenti in un sottogruppo di geni, in modo da ottenere un qualche tipo di informazione che permetta di elaborare dei criteri sulla base dei quali definire delle differenze tra un insieme di pazienti malati ed un altro a cui appartengono solamente individui sani, effettuando un confronto tra le sequenze che compongono i rispettivi acidi nucleici.

Questo permetterebbe di indagare sugli effetti collaborativi che si verificano tra i geni che compongono i vari complessi proteici, nell'attuazione di tutte quelle attività biochimiche che compiono le funzioni cellulari di un qualsiasi organismo vivente.

Sono da considerarsi "effetti collaborativi" tutte quelle interazioni geniche che concorrono, quindi, allo svolgimento della medesima azione, partendo dal presupposto che i geni che costituiscono un determinato complesso proteico non devono essere analizzati come se fossero entità completamente indipendenti le une dalle altre, ma come un insieme di interconnessioni messe in atto da un meccanismo di stretta cooperazione che realizzi sinergia [58].

Assumendo questa visione, allora, risulta evidente che ai fini di comprendere il ruolo che le interdipendenze geniche ricoprano nel manifestarsi di disfunzioni cellulari e malattie, come l'insorgenza e la diffusione di una massa tumorale, si debba analizzare la struttura che caratterizza questi meccanismi cooperativi, per valutare come questi condizionino e si manifestino nel fenotipo di un organismo.

Inoltre, integrando questi dati con le informazioni ricavate dalla misurazione del livello di espressione con cui i geni si presentano e dallo studio di polimorfismi a singolo nucleotide, ovvero delle mutazioni genetiche che interessano le sequenze di un solo nucleotide, tali per cui la variazione si verifica nel profilo genomico di almeno l'1% della popolazione esaminata, sarà possibile ottenere un quadro più completo e generale per quanto concerne la struttura dei meccanismi biologici che interessa una patologia di interesse.

4.3.1 Formulazione del pattern ricercato

Considerano un insieme di SR-networks che modelli il materiale genetico del database N da analizzare, quello che si vuole vefircare è se in questi grafi sia presente una sottostruttura dalle specifiche caratteristiche, che corrisponde di cui si vuole indentificare la ricorrenza.

Per pattern P si intende, quindi, un sottografo connesso che dimostri di avere una corrispondenza nell'insieme di nodi che compongono le reti di tipo Strength-Relevance che rappresentano la popolazione N, sia che questa sia costituita soltanto da individui sani, sia che questa riguardi solamente i pazienti malati.

Questo significa che, nelle SR-networks assegnate, definite come insieme di nodi e archi N = (V, E), sia stato possibile trovare un riscontro del pattern P = (V_p, E_p) per cui valga che i suoi vertici siano contenuti all'insieme dei nodi di N, secondo la relazione $V_p \subseteq V$ e che l'insieme degli archi che esplicano le relazioni tra i geni sia un sottoinsieme degli archi di N, tale per cui $E_p \subseteq E$.

Inoltre, deve essere rispettata la proprietà per cui dato uno specifico pattern P, la cui struttura sia stata identificata in N, allora la corrispondenza trovata può essere una soltanto.

L'algoritmo utilizzato si occuperà, dunque, di ricercare, quel tipo di *pattern* definito come *discriminante*, sulla base del contenuto informativo che presenta [10].

Il cosiddetto guadagno di informazione sarà valutato relativamente alla misurazione dell'entropia [59], ovvero il "disordine" mostrasto dal dataset

utilizzato per la costruizione dei grafi, causata dalla presenza di quello specifico pattern.

Tenendo conto di dover effettuare questa ricerca su delle reti biologiche che rappresentano delle interazioni geniche e gli effetti collaborativi che ne conseguono, si definisce, come pattern discriminante, un determinanto pattern P che, considerato un qualunque suo sottografo P', dimostri di essere comunque caratterizzato da un guadagno informativo maggiore di quest'ultimo.

In conclusione, il potere discriminante del pattern P, messo a confronto con quello della sua sottostruttura P', sarà più significativo di quest'ultimo e descritto dalla seguente relazione, per cui vale che $pow(P) > pow(P')$.

4.4 Fase di Preprocessing dei dati

Affinchè l'algoritmo possa essere applicato, è necessario che i dati della popolazione su cui deve effettuare le operazioni computazionali richieste siano presentati in maniera adeguata.

Più precisamente, è richiesto che i campioni del materiale genico che si vuole analizzare siano rielaborati attraverso una fase di preparazione, a conclusione della quale si dovranno ottenere due differenti dataset, in modo tale che i dati genici siano distinti tra quelli che appartengono ai soggetti affetti dalla patologia di interesse e quelli relativi ad individui riscontrati come sani.

Questo tipo di suddivisione condurrà, quindi, alla generazione di un tipo di dato che descriva le caratteristiche genomiche dei pazienti malati, identificato come dataset "Unhealty" ed un altro che, invece, conterrà tutte le nozioni di cui si dispone sui pazienti sani, denominato come "Healty" set.

4.4.1 Dati : origine e caratteristiche

Il materiale genetico, che è stato oggetto dell'analisi attuata tramite il software sopracitato, è stato ottenuto tramite l'accesso ad una delle banche dati principali in questo ambito, GEO DataSets.

Questo database si occupa della memorizzazione di informazioni relative all'espressione genica, la cui fonte è il repository di Gene Expression Omnibus (GEO), fornendo oltre che i records dei campioni di acidi nucleici, degli strumenti aggiuntivi che permettano di effettuare, più agevolmente, la ricerca di esperimenti e campioni specifici.

Immagine 25: GEO è un repository di dati sull'espressione genica.

I datasets che sono stati utilizzati per l'analisi appartengono al corredo genetico di diversi individui, i quali, attraverso degli esami medici, sono stati dichiarati come soggetti sani oppure riscontrati come affetti dalla una determinata malattia.

La differenziazione dei campioni in due catetegorie distinte è necessaria per fare in modo che l'algoritmo evidenzi, tramite la ricerca di strutture ricorrenti, quali geni siano rilevanti nel manifestarsi o meno della patologia di interesse.

I GEO data impiegati per la sperimentazione di questo elaborato sono stati tutti estratti dalle sequenze nucletiche appartenenti ad organismi di specie Homo Sapiens, perciò si tratta di campioni di soli esseri umani.

Questi quattro datasets verranno descritti qui di seguito:

- "GSE161134", che corrisponde alle informazioni geniche di 310 pazienti, tra i quali è noto che 120 si siano sottoposti a cure chirurgiche poichè affetti da paradontite o piorrea;
- "GSE25724", cioè dei profili di espressione genici ottenuti utlizzando la tecnica dei microarray, riguardanti 7 individui sani e 6 che, invece, risultano essere affetti da diabete di tipo 2;
- "GSE55200", che consiste in una raccolta di dati relativi all'espressione genica, tratti dall'analisi di tessuto adiposo sottocutaneo, che riguardano la condizione di obesità. Tra i piazienti analizzati sono presenti dei soggetti obesi che accusano dei disturbi metabolici e degli individui obesi che possono, però, essere riconosciuti sani dal punto di vista metabolico;
- "GSE68907", un dataset che, aggregando i dati relativi all'espressione genica sulle caratteristiche comuni riscontate, patologicamente, nel cancro alla prostata, si propone di poter identificare dei geni, il cui manifestarsi possa essere utile per predirre il comportamento clinico di questa malattia.

In base al formato in cui questi dati si presentano, saranno sottopposti ad un tipo mirato di azioni, atte a renderli idonei ed utilizzabili dall'algoritmo di cui si è trattato.

Si tratta, in genere, di file con estensione .txt o .soft, che verranno elaborati utilizzando Matlab, un ambiente software per il calcolo numerico e l’analisi statistica.

Per quanto riguarda i dati di interesse, sono tutti descritti sottoforma di file di testo, di tipo “GSExxx.txt”.

Una serie GEO rappresenta un record orginale che è stato inserito nel database come materiale estratto da un insieme di pazienti, per poi essere inserito in una specifica raccolta di campioni, che andrà a costituire un DataSet.

Su questo raggruppamento di campioni sarà possibile effettuare dei confronti biologici e statistici, utilizzando anche degli strumenti della stessa piattaforma, come la visualizzazione tramite dei grafici e la possibilità di clusterizzare degli elementi, secondo dei criteri da stabilire in base alle necessità.

Si considerino, ora, i campioni trattati in questo elaborato e si esplichi il modo in cui è stato necessario rielaborarli ai fini di renderli idonei per l'esecuzione dell'algoritmo.

Per quanto riguarda i datasets "GSE16134", "GSE25724" e "GSE55200", sono stati messi a disposizione sotto forma di serie riportate nei file di testo dai rispettivi nomi e sulla stessa piattaforma dalla quale sono stati prelevati è stato possibile effettuare un'analisi tramite il tool GEO2R, in modo tale da suddividere, già dal principio, i campioni in diversi sottoinsiemi, in base alle esigenze.

In questo caso si è deciso di individuare due soli sottogruppi, definiti con il nome di Healty e Unhealty, in modo da distinguere i pazienti malati dai soggetti che, invece, sono stati riscontrati come sani, assegnando, dunque, gli individui

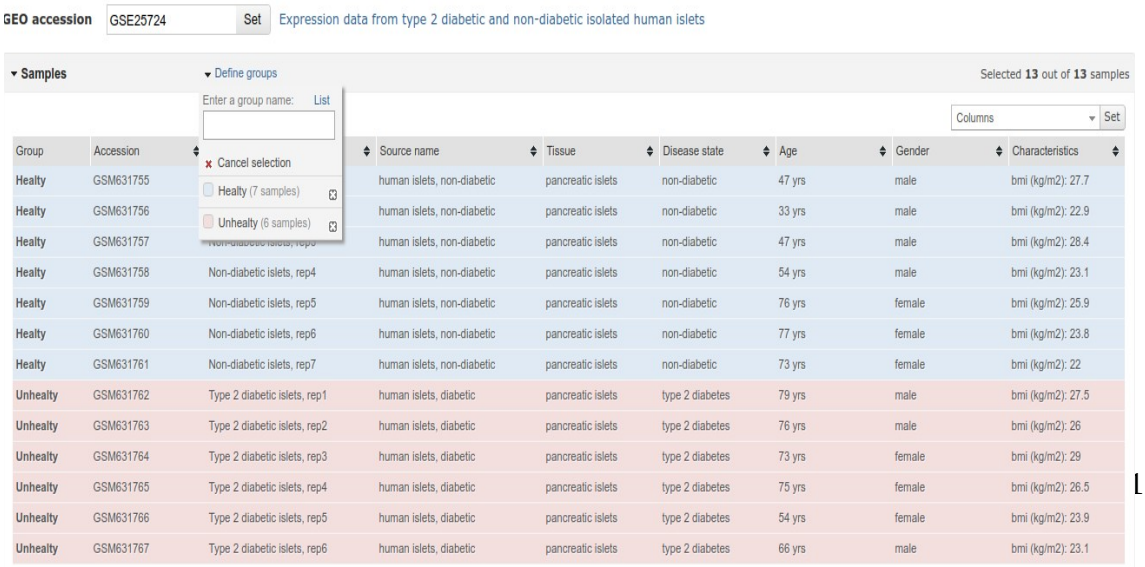


Immagine 26: Esempio di come è stata effettuata l'analisi GEO2R, distinguendo due raggruppamenti di campioni, del dataset GSE25724.

non affetti dalla specifica patologia al primo cluster e gli altri al secondo.

Il file restituito al termine di questo processo includerà tutte le informazioni che si sono ottenute mettendo a confronto le caratteristiche dei campioni appartenenti ai due gruppi differenti, descritte tramite una tabella di geni, ordinati in termini di significatività che hanno mostrato per quanto riguarda il livello di espressione.

Questa azione si è resa necessaria poiché nessuno dei tre insiemi di campioni considerati era stato rielaborato, dalla banca dati in questione, in modo tale da renderlo disponibile come effettivo oggetto DataSet, un tipo di dato organizzato secondo i criteri della piattaforma stessa, che mettesse in risalto le differenze, in termini di manifestazione genica, tra i due gruppi di individui esaminati.

Se si considera il quarto insieme di campioni da utilizzare, denominato “GSE68907”, invece, si può riscontrare come questo sia stato già organizzato e restituito come oggetto DataSet, infatti nel file di testo che lo rappresenta è possibile riscontrare la presenza di valori numerici decimali a precisione singola, che descrivono il livello di espressione di ogni gene individuato, secondo una suddivisione matriciale.

In questo caso,allora, per risalire all’effettivo nome del gene ed al quadro clinico del paziente, è stato necessario utilizzare delle informazioni aggiuntive, rese disponibili sulla piattaforma che si è occupata di fornire lo specifico esperimento, ovvero Platform GPL8300.

Da qui è stato possibile scaricare una tabella completa, salvata come “GPL8300-tbl-1.txt” , contente tutti i nomi dei geni emersi tramite l’esperimento impiegato per la creazione del dataset “GSE68907”, riportati secondo la notazione standard adottata finora, ovvero idGene_at.

Ai fini di recuperare anche i dati che certifichino lo stato di infezione o meno

dei pazienti esaminati, si deve includere, in questa fase di preparazione dei dati, anche la lavorazione del file "GSE68907_family.xml", poiché contiene proprio le informazioni su quali soggetti siano sani e quali malati.

Segue una spiegazione del modo in cui è stata svolta questa fase di preprocessing, che si è occupata dell'elaborazione di tutti i campioni che si è scelto di sottoporre all'algoritmo.

4.4.2 Matlab: elaborazione dei dati

Il software scelto utilizzato per lavorare i GEODatasets è Matlab, poiché provvisto di funzioni in grado di interfacciarsi con questa particolare tipologia di dati.

Siccome, come si è già detto, i dati che devo essere rielaborati richiedono di un approccio differente, in base alla struttura in cui sono descritte le informazioni necessarie, sono state impiegate due funzioni distinte.

Ciò che accomuna questi scripts, ovvero dei codici scritti in Matlab, memorizzati con estensione .m, che possono essere eseguiti direttamente nel suddetto software, riguarda l'insieme dei parametri che ricevono.

Si consideri, dunque, il dataset GSE68907, i cui dati sono forniti sottoforma matriciale.

Qui di seguito si riporta il codice dello script utilizzato per l'estrazione degli elementi di interesse, denominato "prepareDataFromMatrixFileTXT.m" :

```
function[healthy,unhealthy] =
prepareDataFromMatrixTXT(fileName,state,sampleSize,fileGen
esNames)
%dimensione della matrice [righe,colonne]
sizeMatrix = [12625 102];
```

```
%prendo il nome del file (ricevuto come parametro)
%genero il file id leggendo il file da analizzare
fileid = fopen(fileName,'r');
%ho un vettore colonna che contiene tutti i valori scritti
nel file
profileMatrix = fscanf(fileid,'%f',sizeMatrix);
%cosi' viene generata la matrice che si puo' elaborare in
%base al vettore degli stati
%nGenes e' il numero di righe della matrice che io ho
%costruito,cioe' 12625
nGenes=sizeMatrix(1);
%sampleSize e' il numero di geni con cui si vuole lavorare
%e viene specificato come parametro
seed = 27623;
%campionamento per l'estrazione dei geni
rng(seed);
sampledGenes = randperm(nGenes,sampleSize);
%l'estrazione dei nomi dei geni verra' effettuata
elaborando %il file
%in formato txt, che contiene i nomi dei geni presenti nel
%dataset
%codificati in base a due convenzioni(si sceglie xxxxx_at)
%il file di interesse poi viene importato e memorizzato
come %vettore riga
genesNames = importdata(fileGenesNames);
%I nomi dei geni campionati verranno memorizzati
%in un file usando questa funzione
%dopo che il file che li contiene e' stato convertito in
%vettore
getGenesNames(genesNames,sampledGenes,fileName);
%dalla profile matrix, in base al vettore deglie
stati,divido
%il set dei sani ed il set dei malati
healthy = profileMatrix(sampledGenes,state==1);
unhealthy = profileMatrix(sampledGenes,state==0);
%creo i due file
[pathstr,name,~] = fileparts(fileName);
if(isempty(pathstr))
    pathstr = '.';
```

```
end
fileh = [pathstr filesep name '_H.ds2'];
fileu = [pathstr filesep name '_U.ds2'];
healthy = prepare(healthy);
unhealthy = prepare(unhealthy);
dssave(fileh, healthy);
dssave(fileu, unhealthy);
end
%funzione prepare
function[profileShift] = prepare(matrix)
    % vettore delle medie
    meanVector = mean(matrix,2);
    %vettore colonna delle deviazioni standard
    ds = std(matrix,0,2);
    profileShift = (matrix -
meanVector*ones(1,size(matrix,2)))./(ds*ones(1,size(matrix
,2)));
end
```

Per poter eseguire questo script occorre che si abbiano a disposizione:

- il path o percorso in cui il file da analizzare è situato nel file system;
- il percorso del file di testo in cui sono elencati i nomi dei geni coinvolti nei profili genici dei pazienti, denominato “GenesList.txt” ;
- un vettore binario, che consiste nel parametro “state”;
- il parametro “sampleSize”, che corrisponde al numero di geni che si vuole siano estratti da tutta la lista di cui si dispone, per generare le informazioni sulla base delle quali creare, in seguito, la rete dei pazienti malati e quella dei pazienti sani.

L’insieme di dati “GSE68907” è risultato quello più problematico da gestire durante la fase di preprocessing, perché essendo stato rielaborato dalla piattaforma, per risalire a tutte le informazioni di interesse, è necessario che si effettuino delle determinate azioni, in modo da ricavare tutti gli elementi

necessari allo script matlab che è stato scritto ad-hoc, proprio per la gestione di una tipologia di dati come questo.

Per prima cosa si devono recuperare i codici id dei geni, cioè i loro nomi identificativi, utilizzando il file “GPL8300-tbl-1.txt”, come è già stato anticipato.

A questo scopo si è pensato di scrivere un metodo, in linguaggio Java, che se ne occupasse, denominato “analisysGenesFile”, di cui si riporta il codice:

```
    private static void analisysGenesFile(File fgenes) {

        //metodo che restituisce un file con tutti i
nomi            //dei geni
        controlloFile(fgenes);
        //creo il nome del file dei risultati
        String fgenesName = new
String("[path]/analyzedSamples/GSE68907/GenesList.txt");
        //creo il file dei risultati
        File geneNames = new File(fgenesName);
        //per contare i geni
        int ngenes=0;
        //creo uno stringbuilder che uso per mantenere i
            //dati nel
        //file di output
        StringBuilder sb = new StringBuilder();
        //uso lo string tokenizer per elaborare il file
di      //testo
        //di ogni riga devo prendere solo la prima
stringa
        //si deve leggere il file, si usa la classe
scanner
        try {
            Scanner scf = new Scanner(fgenes);
            while(scf.hasNext()){
                //si preleva la riga
                String line = scf.nextLine();
                //si deve processare la prima riga
                StringTokenizer st = new

    StringTokenizer(line);
```

```
                    //in modo da estrarre solo la prima
stringa
                    String token = st.nextToken();
                    if(token!=null){
                         //questo token e' valido
                         sb.append(token);
                         sb.append('\n');
                         ngenes++;
                    }
               }//while
               //si chiude il flusso in lettura
               scf.close();
          }catch (FileNotFoundException e) {
               System.err.println("errore in lettura");
          }
          //l'intero contenuto dello stringBuilder deve
essere         //scritto su file
          //prima lo converto in vettore di stringhe
diviso              //per righe
          //usando come separatore il fine riga
          String [] contenuto = sb.toString().split("\\
n");
          try {
               //il costruttore riceve il file in cui
scrivere
               PrintWriter pf = new PrintWriter(geneNames);
               //itero il vettore di stringhe per righe
               for(String s : contenuto) {
                    //scrivo la riga su file
                    pf.append(s);
                    pf.append("\n");
               }
               //si svuota il buffer
               pf.flush();
          } catch (FileNotFoundException e) {
                    System.err.println("errore in
scrittura");
          }
```

```
    }//analysisGenesFile
```

Questa funzione si occupa di iterare il file ricevuto come parametro, per estrarre soltanto le stringhe che contengono i nomi dei geni, una per ogni riga del documento.

L'output sarà, quindi, un file di testo che contiene tutti i nomi facenti parte delle sequenze nucleotidiche appartenenti all'esperimento.

Il passo successivo è la creazione del vettore binario degli stati, composto cioè da elementi che possono essere soltanto numeri pari a 0 oppure ad 1.

Lo scopo è quello di certificare se il paziente in esame sia malato, associando ad esso il valore 1 (Unhealty), oppure che non risulti affetto dalla patologia esaminata, identificandolo con il numero 0 (Healty).

Per poter istanziare un vettore di questo tipo si deve essere a conoscenza, quindi, del quadro clinico dei pazienti e dell'ordine in cui i loro campioni siano stati riportati nel GEODataSet.

Questo elenco è riportato nel file "GSE68907_family.xml", fornito anch'esso dalla piattaforma, che contiene una serie di informazioni relative all'esperimento che è stato condotto.

Per estrarre soltanto quelle di interesse, ovvero quali tra i soggetti analizzati sia malati e quali sani, è stato creato un metodo, scritto in linguaggio Java, che se ne occupasse, restituendo, come output, il vettore degli stati.

Questo metodo è denominato "analisysXMLFile" ed è una funzione realizzata dalla classe "FileDataAnalysis.java", che realizza anche quella descritta sopra, dedita all'estrazione dei nomi dei geni.

Segue il codice del metodo:

```
    //metodo che si occupa dell'analisi del file
    private static void analysisXMLFile(int
nsamples,String        fname) throws SAXException,
```

```
IOException,
     ParserConfigurationException{
          //questo metodo riceve il nome del file da
               //analizzare
          if(fname==null)throw new
                    IllegalArgumentException("parametro
nullo");
          //vettore binario degli stati da restituire
          //stampato su file
          int[] state = new int[nsamples];
          //si accede al file da utilizzare
          File f = new File(fname);

          controlloFile(f);

          //il vettore degli stati verra' generato e
stampato       //su file
          File vstate = new
               File("[path]/Software/analyzedSamples/"
                         + "GSE68907/vstate.txt");
          //per istanziare il parser che si occupera'
               //dell'analisi
          //sintattica uso il Document Object Model (DOM)
          //istanzio il metodo factory dell'oggetto
document       //da elaborare
          DocumentBuilderFactory dbf =

     DocumentBuilderFactory.newInstance();
          //istanzio il builder per creare il file
document
          DocumentBuilder db = dbf.newDocumentBuilder();
          //creo un document in modo da fare il parsing
sul            //file
          Document d = db.parse(f);
          d.getDocumentElement().normalize();
          //estraggo il primo nodo
          Node root = d.getDocumentElement();
         //estraggo tutti i nodi della lista
```

```
caratterizzati              //dal tag "<Source>"
        NodeList nList =
d.getElementsByTagName("Source");
        //in questo modo ottengo, nell'ordine,
l'informazione        //che sara'o "normal" o "tumor",
utile per generare          //il vettore degli stati
        for (int i = 0; i < nList.getLength(); i++) {
             //prendo il nodo corrente
             Node nCurr = nList.item(i);
             //prendo le informazioni che contiene
             Element e = (Element)nCurr;
             //stampo il contenuto all'interno del tag
             //devo costruire un vettore binario di 102
                  //celle che contiene 1 se il paziente
e' sano,          //ovvero se la stringa stampata e'
normal, 0                  //altrimenti, cioe' se la
stringa stampata e'                  //tumor
             String s = e.getTextContent();
             if(s.equals("normal")) {
             //allora nel vettore metto 1
                  state[i]=1;
             }
             else {
                  //allora il tag e' tumor
                  //nel vettore metto 0
                  state[i]=0;
             }
             //con l'incrementare di i del for, scorre
         //anche l'indice del vettore
        }//for
        //stampo i risultati su file
        stampaVettore(state,vstate);
    }//analysisXMLFile
```

Allora, questa funzione, sfruttando gli strumenti forniti dal linguaggio Java proprio per gestire file nel formato .xml, va a ricercare il contenuto racchiuso nel tag *<Source>[normal/tumor]</Source>*, indicativo dello stato clinico

del paziente, in un documento di testo sì fatto:

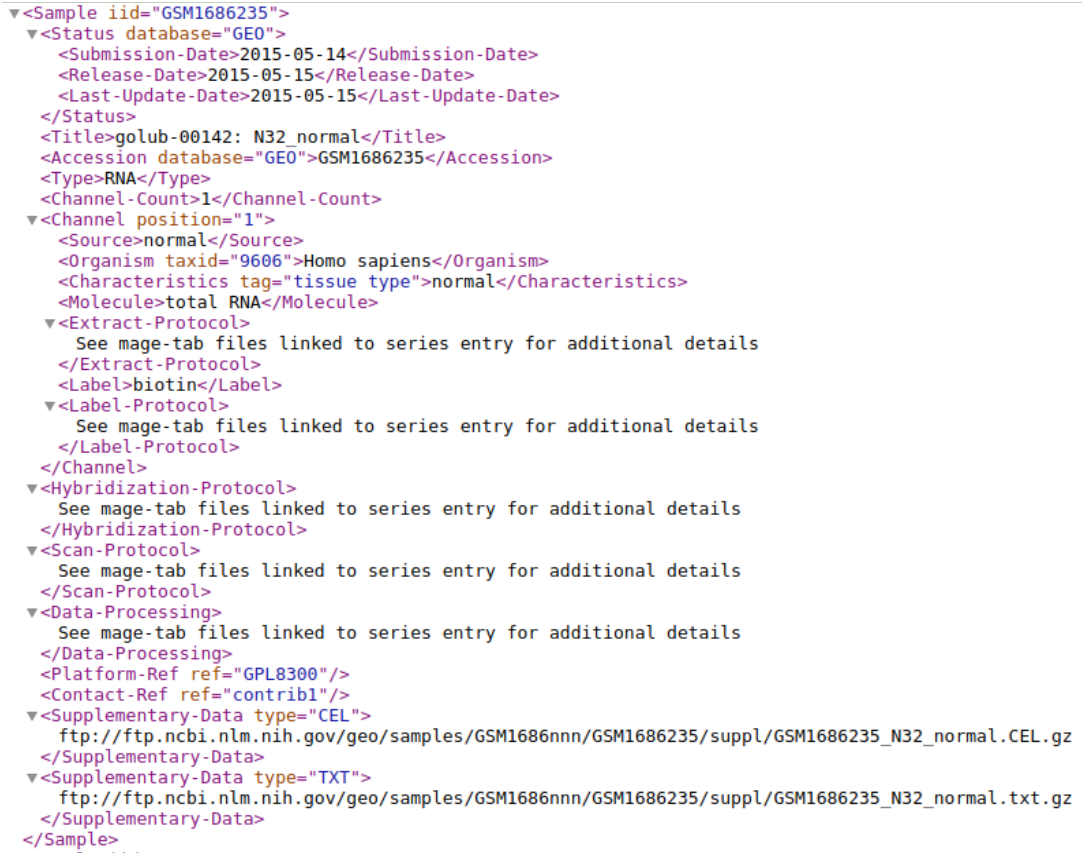


Immagine 27: Blocco di codice del documento .xml che descrive le informazioni relative ad uno specifico campione, con identificativo "GSM1686235".

A questo punto può essere eseguito lo script matlab, che riceverà i parametri elencati in precedenza ed effettuerà l'analisi di questo GEODataSet in base al valore di sampleSize specificato.

Per quanto riguarda gli altri sets di campioni, il processo di elaborazione che è stato utilizzato è simile a quello adottato per il dataset GSE68907, anche in questo caso si richiede un numero inferiore di azioni da effettuare, al fine di renderli adatti per eseguire l'elaborazione.

Infatti, siccome i tre insiemi di campioni da analizzare, ovvero GSE16134,

GSE25724 e GSE55200, sono stati forniti dalla piattaforma come documenti di testo contenenti “serie” di dati, è sufficiente che vengano soltanto eleborati tramite uno script, simile al precedente, denominato “prepareDataFromFileTXT.m”.

In questo caso, potendo analizzare i campioni con il tool “GEO2R”, questi verranno suddivisi in due gruppi, in base alle categorie specificate, ovvero Healty ed Unhealty e da questo elenco è possibile ricavare il vettore degli stati.

Non è necessario estrarre anche l’elenco dei nomi dei geni, in quanto questi sono inclusi nel file che descrive ciascun set di pazienti.

Per automatizzare il processo di elaborazione dei dati su cui si fonda questa fase di preprocessing è stato creato un ulteriore script, che si occupasse di eseguire tutte le chiamate ai rispettivi scripts che gestiscono le analisi effettive dei quattro datasets GSExxx.

Al suo interno è stato anche definito l’intervallo in cui si vuole che vari il parametro sampleSize, cioè il numero di geni da campionare, in maniena casuale, tra gli agglomerati proteici che costituiscono i profili genici dei pazienti.

In base alle necessità, dunque, questo dovrà essere specificato, prima del lancio dello script denominato “gsedataAnalysis.m”, all’interno dello stesso.

Si riporta il codice in questione, in cui si evince anche che si è scelto di lavorare con un numero di geni pari a 100:

```
%il parametro sampleSizeche variera' nell'intervallo
seguente:
sampleSizeArray = [10 50 100 150 200];
%valore attuale
%sampleSize = sampleSizeArray(1);
%sampleSize = sampleSizeArray(2);
sampleSize = sampleSizeArray(3);
```

```
%sampleSize = sampleSizeArray(4);
%sampleSize = sampleSizeArray(5);

%GSE 16134
fileName16134 = '[path]/GSE16134.txt';
%vettore degli stati di GSE 16134
state16134 = [ 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0 0 1 0
0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 0
0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0
1 0 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0
0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1  0 0 0 0 0 0 0 0
1 0  0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0
0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ];
%chiamata alla funzione al variare di sampleSize
prepareDataFromFileTXT(fileName16134,state16134,sampleSize
);

%GSE 25724
fileName25724 = '[path]/GSE25724.txt';
%vettore degli stati di GSE 25724
state25724 = [ 1 1 1 1 1 1 1 0 0 0 0 0 0 ];
%chiamata alla funzione al variare di sampleSize
prepareDataFromFileTXT(fileName25724,state25724,sampleSize
);

%GSE 55200
fileName55200 = '[path]/GSE55200.txt';
%vettore degli stati di GSE 55200
state55200 = [ 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0
0 ];
%chiamata alla funzione al variare di sampleSize
prepareDataFromFileTXT(fileName55200,state55200,sampleSize
);
```

```
%GSE 68907
%non ho il file nel formato dei precedenti
%quindi per analizzarlo devo usare un'altra funzione
fileName68907 = '[path]/GSE68907.txt';
%vettore degli stati di GSE 68907
state68907 = [ 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 0 1 1 0 0 0
1 0 0 1 1 1 1 1 0 0 1 1 0 0 1 0 1 0 1 1 1 0 0 0 1 0 0 1 0
0 0 0 0 1 0 1 0 0 1 1 1 1 0 0 0 0 0 1 0 1 0 0 0 0 1 1 0 0
1 1 0 0 1 1 1 0 1 1 0 1 1 0 0 0 1 0 1 0 0 0 ];
%file che contiene i nomi dei geni elaborato tramite java
fileGenesNames = '[path]/GSE68907/GenesList.txt';
%chiamata della funzione al variare di sampleSize
prepareDataFromMatrixTXT(fileName68907,state68907,sampleSi
ze,fileGenesNames);
%end
```

4.4.3 Output della fase di preprocessing

A conclusione della fase di preprocessing dei dati, ci si aspetta che vengano generati tre differenti file, per ognuno dei datasets rielaborati, ovvero:

- un file di testo, dal nome “**sampledGenes_GSExxx.txt”**, che contiene l’elenco dei nomi dei geni coinvolti nell'esperimento.
  Nello specifico, il nome il nome riportato sulla riga i-esima corrisponde al gene presente sulla riga i-esima della matrice che è stata costruita;

- **GSEXXXX_H.ds2**, che rappresenta una matrice del tipo (nGenes x nSamples) contenente i dati relativi ai campioni sani;

- **GSEXXXX_U.ds2**, cioè la matrice di dimensione (nGenes x nSamples) che descrive i dati che riguardano i campioni dei pazienti identificati come malati;

Ogni campione di questi set di dati elaborati rappresenta il *trascrittoma* del paziente analizzato, cioè il suo patrimonio genetico.

Le matrici ottenute, che sono state memorizzate nella stessa directory in cui è stato eseguito lo script di analisi dei GEODataSets, saranno costruite secondo la seguente struttura:

- ogni riga della matrice rappresenterà lo specifico gene, estratto in modo random tra quelli presenti nei trascrittomi dell'esperimento;
- ogni colonna della matrice corrisponderà al paziente su cui si intende ricercare la presenza del suddetto gene;

Perciò, ogni cella della matrice conterrà la rilevanza dello specifico gene, in quel determinato campione.
In conclusione, questi sono i dati che andranno a costituire l'input della parte del software che si occupa di costruire i grafi da associare alle informazioni ricavate proprio in questa fase di preprocessing.
Più precisamente, si andrà a generare una coppia di grafi da associare a ciascuno dei datasets presi in esame, in modo da rappresentare distintamente l'insieme dei soggetti malati e degli individui sani.

4.5 Analisi dei campioni e risultati

Si procede ora con la descrizione della parte svolta effettivamente dal software, distinta in due parti principali.

Nella prima si procederà con la creazione dei grafi che rappresentano i campioni processati nella fase precedente, per poi analizzarli, nella seconda, in modo tale da ricercare delle sottostrutture discriminanti, nell'una e nell'altra

rete che costituiscono proprio l'input dell'algoritmo, ovvero quella associata ai pazienti sani e quella associata ai pazienti malati.

4.5.1 Costruzione delle reti

Per generare i grafi, sui quali poi verrà applicata la tecnica di pattern mining che si è statibilito di adottare, è necessario utilizzare i due file identificati, rispettivamente, come GSExxx_H.ds2 e GSExxx_U.ds2, che sono stati preparati durante la fase di preprocessing.
Siccome questa parte di software utilizza delle funzioni scritte in linguaggio C, sono stati definiti due scripts in bash, che permettessero di automatizzare la procedura di creazione dei grafi, atta a modellare la coppia di reti Healty ed Unhealty da impiegare per la descrizione di ogni dataset.
Ciascuno dei due script si occupa, in particolare, di invocare la funzione dedita alla costruzione della rete, sulla base dei valori che descrivono le caratteristiche che si vuole abbiano i due grafi, ovvero:

- quali geni andranno a costituire i nodi di ciascuna;
- quali valori si vuole che si adottino come soglie, o treshold, sulla base dei quali stabilire se esista o meno un arco tra una coppia di nodi, formata da due geni;

Quello che si otterrà al termine di questa procedura saranno, dunque, due grafi, ciascuno dei quali sarà rappresentato tramite una matrice di adiacenza.
Si ricordi che gli archi dei grafi saranno provvisti di pesi, misurati in termini di robustezza τ_s e rilevanza τ_r .

4.5.2 Discriminative Patter Mining sui Grafi

Avendo a disposizione due grafi H e U, associati alle sottopopolazioni, rispettivamente, di pazienti sani e malati, in cui sono stati suddivisi i GEODataSets, è possibile avviare l'algoritmo che implementa la tecnica di pattern mining che si deve adottare, ovvero quella che ha come scopo la ricerca di sottoreti che siano discriminanti.

Anche in questo caso, per invocare la funzione che implementa l'algoritmo di mining, sono stati utilizzati due scripts, scritti sempre in Bash, in modo tale da ripetere agevolmente i due tipi di analisi dei campioni, ovvero quella attuata al variare del numero di geni coinvolti nelle struttura delle reti e quella che, invece,

esegue la ricerca valutando la significatività dei patterns in base ai valori di threshold modificati ogni volta.

Quello che si genererà in output sarà un file di testo che conterrà il risultato ottenuto come conclusione della ricerca di patterns discriminanti, all'interno dell'una e dell'altra rete.

Nello specifico, le informazioni riportate saranno relative a:

- i parametri utilizzati come termini da rispettare nella ricerca, siano questi il numero di geni scelto per la costruzione delle reti ed i due valori di soglia, ovvero Strength Threashold e Relevance Threshold, ma anche il numero di campioni esaminati;
- l'elenco degli archi che compongono il grafo, sia esso quello campioni dei sani H o quello dei malati U, in cui ogni arco viene descritto riportando la coppia di geni tra cui è stato tracciato e la significatività che lo caratterizza, misurata in termini di score ed entropia, calcolata per la specifica relazione;

- un elenco di patterns che si è riscontrato fossero superiori alla soglia fissata;
- un ulteriore elenco di patterns, identificato come top-50, che riporta i 50 patterns che sono stati identificati come i più discriminanti;
- il numero di Visited Nodes, ovvero il numero di patterns che l'algoritmo ha dovuto analizzare nel complesso, in ciascuna delle due reti;

Quest'ultimo rappresenta il valore di interesse sulla base del quale si è deciso di effettuare la sperimentazione, ai fini di valutare la sua dipendenza dai parametri sopraindicati, ovvero le due soglie e il numero di geni.

Vista l'esigenza manifestata dai biologici di dover idenficare quali siano i geni coinvolti in queste relazioni di coespressione, le quali si sono mostrate essere significative ai fini di comprendere la correlazione che sussiste tra il genotipo ed il fenotipo e considerando il contributo che l'analisi del trascrittoma di un organismo possa apportare ai fini di prevedere la possibilità che si verifichi o meno l'insorgenza di una specifica malattia o disfunzione cellulare, si è deciso di introdurre nel software utilizzato uno strumento che soddisfacesse queste premesse.

Allora, si è creata una classe di utilità, scritta in java, denominata "IdentifyPatternGenesName", il cui cuore è rappresentato dalla funzione "**matchingGenes**", invocabile direttamente dal programma di esecuzione dell'algoritmo di *Discriminative Pattern Mining*.

Quello che si propone di effettuare è elaborare il file dato in output dall'algoritmo di mining, per poter identificare quali siano effettivamente i nomi dei geni che costituiscono gli archi dei patterns che sono stati riportati come tra i più discriminanti.

Segue il codice che realizza quanto preannunciato:

```
public final class IdentifyPatternGenesName {

     //costruttore privato
     private IdentifyPatternGenesName() {}
     //array di stringhe che contiene i nomi dei geni
estratti
     //tramite matlab dall'analisi dei dataset dei
pazienti
     private static String [] geneNameArray;
     //la dimensione dell'array viene impostata usando il
     //parametro numNodi =
     //util.OpenBinaryFiles.open(path_h+"/corr0_H.ds2");
     //come numero di geni uso un valore di default
     private static int genesNumber= 200;
     //file di output
     private static File genesNameList;
     //nome di default del file di output
     private static String res_name =
                                   "DiscriminativePatternGen
esName";
     //per le stampe di debugg
     private static boolean debugg = false;
     //funzione che controlla la validita' del file
     private static void analyzingFile(String path) {
          //controllo se il path del file e' nullo
          if(path==null)throw new
               IllegalArgumentException("Path del file
nullo");
          //si crea un'istanza di file dal path, che
contiene      //anche il nome, ricevuto come parametro
          File f = new File(path);
          //controllo se il file esiste e se ho i permessi
di        //lettura
          if(!f.exists()||!f.canRead()){
               try {
                    throw new FileNotFoundException();
```

```
            } catch (FileNotFoundException e1) {
                System.err.println("File non trovato o
non
    accessibile");
            }
        }//controllo file
        if(debugg)System.out.println("Il path e' "

    +f.getPath());
        //salvo il contenuto del file da leggere in un
            //vettore di stringhe, in cui ogni cella
rappresenta       //l'indice di riga del file
        //si istanzia il vettore
        geneNameArray = new String[genesNumber];
        //indice per scorrere il vettore
        int index = 0;
        //per leggere il file si usa la classe scanner
        try {
            Scanner scf = new Scanner(f);
            while(scf.hasNext()){
                //si preleva la riga che contiene il
nome                 //del gene
                String s = scf.nextLine();
                //si inserisce il nome del gene nella
cella                 //del vettore
                geneNameArray[index]=new String(s);
                //si aumenta l'indice di inserimento
nel                 //vettore
                index++;
            }//while
            //si chiude il flusso in lettura
            scf.close();
        }catch (FileNotFoundException e) {
            System.err.println("errore in lettura");
        }
        if(debugg)System.out.println("analyzingFile");
    }//analyzingFile
    //avendo il vettore che contiene l'elenco dei nomi
dei      //geni devo iterare gli archi del grafo che
```

```
risultano          //coinvolti nei patters discriminanti e
genero un file          //che contiene queste informazioni
    private static void namingGenes(List<Pattern>
patterns){
         //questo metodo itera la struttura dati che
contiene      //le informazioni sui pattern piu'
discriminanti,              //dove sono specificati gli
archi
         //coinviolti nella relazione e converte il
codice id          //di questi nel nome corrispondente,
mettendo              //l'output su file
         //per scrivere su file si puo' utilizzare la
classe         //PrintWriter
         StringBuilder sb = new StringBuilder();
         //questo conterra' i dati da stampare su file
         //quindi dalla lista dei pattern se ne estrae
uno e          //si descrivono le sue caratteristiche
         int i = 1;//per numerare il pattern in esame
         for(Pattern p : patterns) {
              //pattern analizzato
              sb.append("Pattern #"+i+"\n");
              List<Integer> edges = (List<Integer>)

    p.getPattern();
              //lista degli archi per quel pattern
              //prelevo la matrice di adiacenza associata
al            //dataset analizzato
              int[][] edgesMatrix =
Dataset.getEdgeMapping();
              //si stampa il numero di archi trovati per
il                 //pattern in esame
              sb.append("Num. archi: "+edges.size()+" ");
              sb.append('[');
              for(Integer e: edges){//itero gli archi
                   //ogni arco e' composto da due geni
                   //per sapere quali devo accedere alla
                        //matrice di adiacenza
                   int g1 = edgesMatrix[0][e];
```

```
                    int g2 = edgesMatrix[1][e];
                    //i due id ottenuti devo convertirli in
                         //nomi accedendo al vettore dei
nomi creato                   //precedentemente
                    sb.append("<"+geneNameArray[g1]+","

     +geneNameArray[g2]+">"+",");
                    //l'arco costituito dai nomi dei due
geni                      //e' stato inserito
               }
               //ho iterato tutti gli archi per quel
pattern
               //per non avere la virgola come ultimo
elemento
               sb.setCharAt(sb.length()-1,']');
               i++;//indice del pattern
               sb.append("\n");
          }//itero i patter
          if(debugg)System.out.println("contenuto\n"+sb);
          //l'intero contenuto dello stringBuilder deve
essere         //scritto su file
          //prima lo converto in vettore di stringhe
diviso              //per righe, usando come separatore il
fine riga
          String [] contenuto = sb.toString().split("\\
n");
          try {
               //il costruttore riceve il file in cui
scrivere
               PrintWriter pf = new

     PrintWriter(genesNameList);
               //itero il vettore di stringhe per righe
               for(String s : contenuto) {
                    //scrivo la riga su file
                    pf.append(s);
                    pf.append("\n");
               }
               if(patterns.isEmpty())pf.append("Non ci sono
```

```
                                        patterns
discriminanti");
                //si svuota il buffer
                pf.flush();
            } catch (FileNotFoundException e) {
                System.err.println("errore in scrittura");
            }
            if(debugg)System.out.println("NamingGenes");
        }//namingGenes

        public static void matchingGenes(List<Pattern>
patterns,               String sup_path, char id,boolean
path_idoneo) {
            //si riceve il parametro id di tipo char
            //che identifica la rete su cui effettuare
l'analisi
            //(verra' usato nel nome del file)
            //si riceve il path del file
            //generato dallo script matlab che contiene i
nomi            //dei geni che compongono i pattern piu'
                    //discriminanti
            //dal path devo estrarre il nome del dataset
            //se il path e' presente nella notazione
standard
            String path="";
            if(path_idoneo) {
                String part= sup_path.substring(0,

        sup_path.lastIndexOf('/'));
                String
geo=part.substring(part.lastIndexOf('/')
                            +1,part.length());
                path=sup_path+"/sampledGenes_"+geo+".txt";
            }else {
                path=new String(sup_path);
            }
            if(debugg)System.out.println(path);
            //si analizza il file
```

```
            analyzingFile(path);
            //si crea il file dei risultati
            createOutputFile(path,id);
            //si effettua l'operazione di matching scrivendo
sul         //file
            namingGenes(patterns);
            if(debugg)System.out.println("Matching
Complete !");
        }//matchingGenes

        public static void createOutputFile(String path,char
id){
            //determino il path in cui verra' istanziato il
file
            //a partire dal path del file da cui ho estratto
i           //dati, ricevuto come parametro
            //seguono delle operazioni sul path
            int last_slash = path.lastIndexOf('/');
            String res_path =
path.substring(0,last_slash+1);
            if(debugg)System.out.println("result path "

        +res_path);
            //determino il nome che dovra' avere il file,
ovvero
            //DiscriminativePatterGenesName_GSEXXXXX.txt
            int name_index = path.lastIndexOf('_');
            String dataset_name =
path.substring(name_index);
            if(debugg)System.out.println("dataset name "

        +dataset_name);
            //creo il nome del file
            res_name +=Character.toString(id)+ dataset_name;
            //creo il file nello stesso path del file da cui
                //posso estrarre i nomi
            genesNameList = new File(res_path+res_name);
            //il file e' stato creato con il nome stabilito
        }//createOutputFile
```

```
}//IdentifyPatterGenesName
```

Nello specifico, questa classe fornisce una funzione che riceve in ingresso, cioè come parametro su cui lavorare, il nome del file, comprensivo di path, da elaborare per poter estrarre i nomi dei geni, ricavati durante la fase di preprocessing tramite le funzioni scritte in Matlab.

Dal punto di vista procedurale, il file in questione viene iterato ai fini di cercare il gene di interesse nella riga che contiene il codice identificativo corrispondente.

Una volta che entrambi i nomi dei geni, trascritti secondo la notazione che si è deciso di adottare nel corso di tutte le operazioni atte ad analizzare i campioni, sono stati selezionati, allora questi verrano inseriti come elementi costituitivi dell'arco che esiste fra i due.

I due file di testo che verranno generati in output, denominati, rispettivamente, "DiscriminativePatternGenesNameH_GSExxx" e "DiscriminativePatternGenesNameU_GSExxx", conterranno, quindi, un elenco degli archi che compongono i patterns classificati come eccezionali, per le reti H e U (rappresentative, rispettivamente, dei pazienti sani e di quelli malati).

In conclusione, per ciascun arco riscontrato verrà riportata la coppia di geni che lo compone, i quali saranno identificati con il loro nome scientifico.

Capitolo 5: Sperimentazione

Per la fase di sperimentazione che è stata realizzata, si è scelto di considerare quattro differenti datasets di campioni genici, già descritti in precedenza, resi accessibili dalla banca dati GEO DataSet.

Questi insiemi di dati, allora, saranno utilizzati per eseguire due principali tipi di test, in modo tale da valutare il comportamento dell'algorimo in base ai differenti parametri che sono stati specificati sia nella fase di preparazione dei dati, sia nel momento in cui è stata avviata l'effettiva analisi degli stessi.

Come si è già detto, i campioni appartengono a pazienti che sono caratterizzati da un diverso quadro clinico, poichè distinti in individui sani e malati, ma accumunati, se appartenenti allo stesso dataset "GSExxx", dall'essere stati sottoposti ad esami medici atti ad estrarre informazioni relative alla stessa patologia di interesse.

5.1 Prima analisi

Dall'analisi dei quattro datasets che si avevano a disposizione, si è ottenuta, per ciascuno di questi, una serie di informazioni, comprensive del numero di patterns che è stato possibile riconoscere all'interno delle reti analizzate.

I grafi su cui l'algortimo ha applicato la tecnica di discriminative pattern mining sono stati costruiti in modo tale che ogni geodataset venisse rielaborato generando due reti, una rappresentante i pazienti malati ed una che descrivesse, invece, i pazienti sani.

Queste coppie di grafi sono state costruite, ogni volta, in base ad un determinato parametro, denominato "sample Size", che è stato specificato nella fase di preprocessing degli stessi dati.

Si ricorda che questo parametro indica il numero di geni sulla base dei quali si aveva interesse che fosse costruita ogni rete, in modo da poter, così, valutare il numero dei patterns che è stato possibile identificare, in correlazione al numero di nodi che costituiscono il grafo esaminato.

Nello specifico, i valori utilizzati per questo parametro variano in un determinato intervallo, che è il seguente:

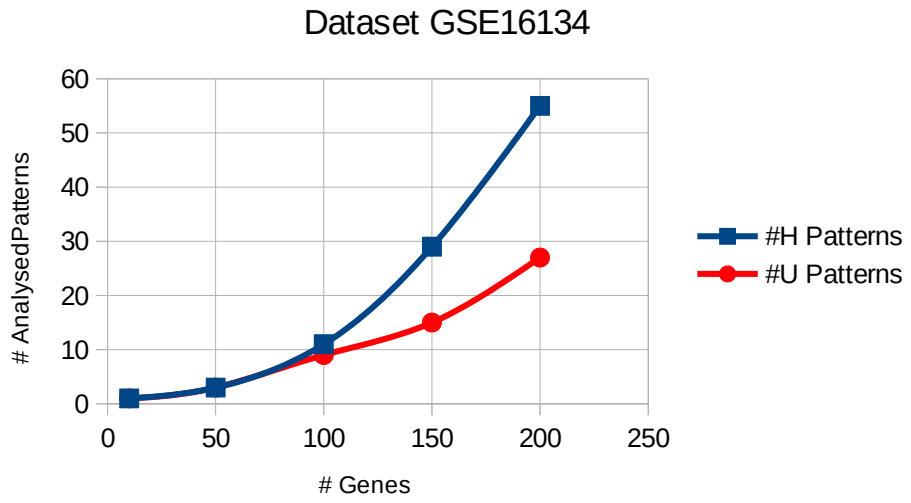
sampeSize = {10, 50, 100, 150, 200}.

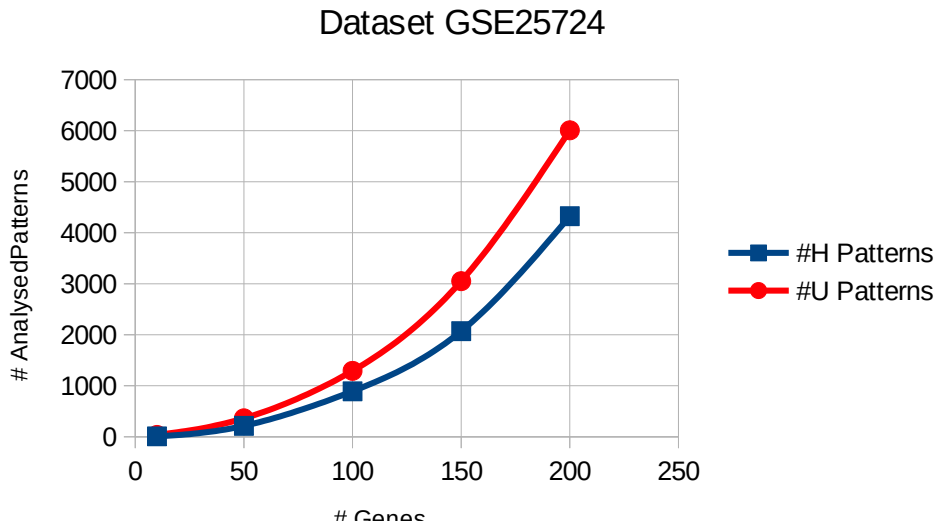
L'analisi che si vuole eseguire ha, come obiettivo, quello di mettere in relazione il numero di nodi di ciascuna delle due reti create, ovvero la rete H associata all'insieme dei pazienti sani e la rete U associata a quello dei pazienti malati, con il numero di patterns che si è riusciti ad identificare, indicato dal parametro "Visited Nodes".

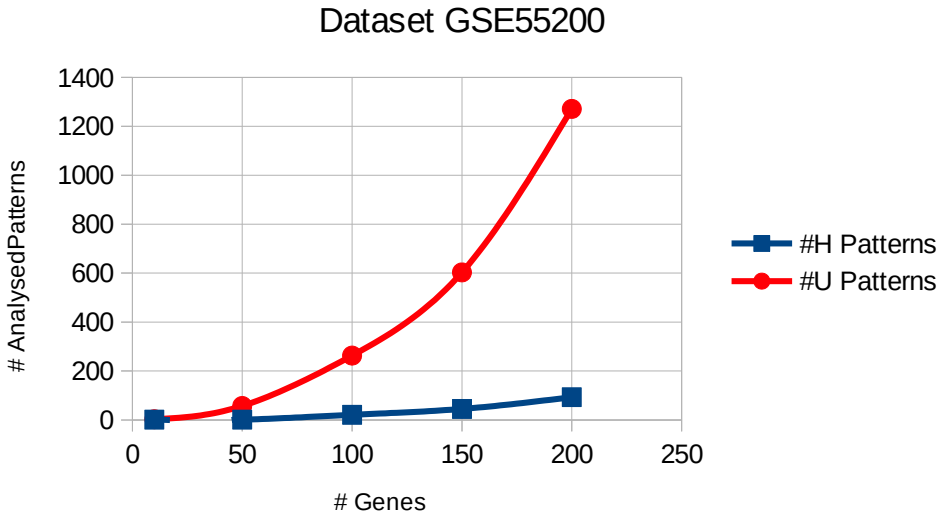
Seguono, allora, i risultati ottenuti per l'analisi di ciascuno dei quattro datasets, corredati da un grafico esplicativo, che metta a confronto il numero di patterns analizzati nella rete dei soggetti sani, con quello calcolato per la rete dei malati.

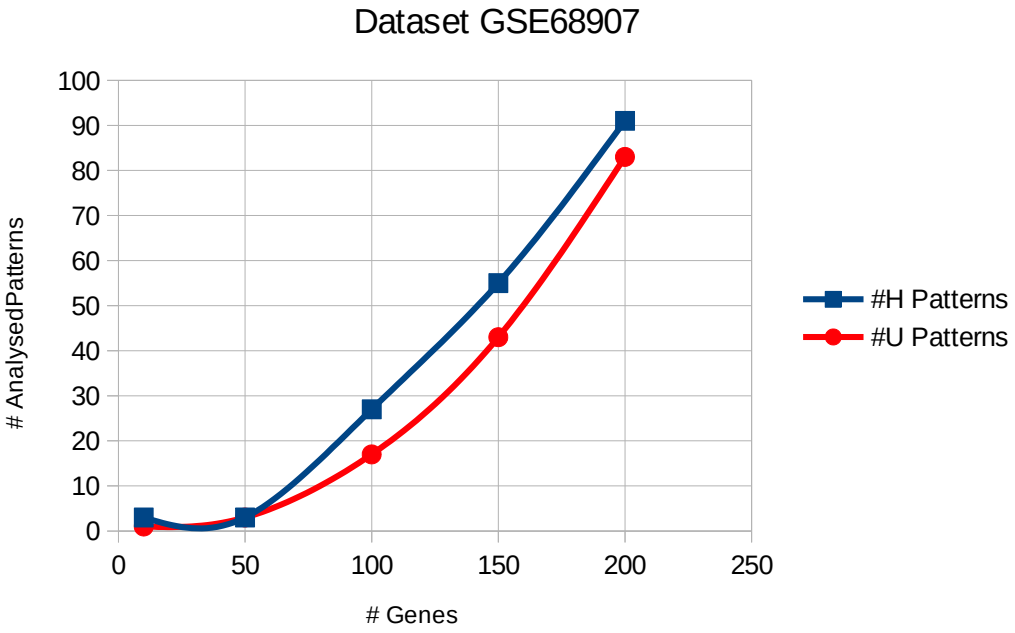
Allora, in ogni grafico sarà possibile osservare due differenti funzioni, denominate, rispettivamente, "H Patterns" se rappresenta il numero di patterns identificati nella rete degli individui sani e "U Patterns" se si riferisce alla rete dei malati, entrambe dipendenti dalla variazione del numero di nodi che le compongono, denominato "# Genes".

Seguono i grafici:









In ogni grafico si è voluto mettere a confronto le funzioni che rappresentano i valori ottenuti per le due reti di ciascun dataset.

Si può notare come il numero di patterns visitati, ovvero quelli analizzati durante la fase di mining, si mantenga comunque inferiore al quadrato degli archi che compongono il grafo di interesse.

5.2 Seconda analisi

L'intento di questa seconda fase della sperimentazione è quello di valutare l'andamento della curva che descrive il numero di patterns analizzati in funzione dei parametri di soglia o threshold, poiché mantenuti fissi a dei valori di default nell'analisi precedente.

Le due reti H e U, rappresentative di ognuno dei quattro datasets in esame, sono state costruite utilizzando un numero di geni pari a 100, essendo un valore mediano considerato il range di valori assumibili dal parametro sampleSize.

Si vedano, nel dettaglio, i parametri coinvolti in questa sperimentazione, ovvero:

- il numero di patterns analizzati, definito come "AnalysedPatterns", come è già stato fatto per il caso precedente;
- il set di valori assumibili dal parametro τ_s , in funzione del parametro τ_r , i quali rappresentano le costanti di threshold, cioè i valori di soglia utilizzati nella creazione dei grafi e per effettuare i tagli durante la fase di mining.

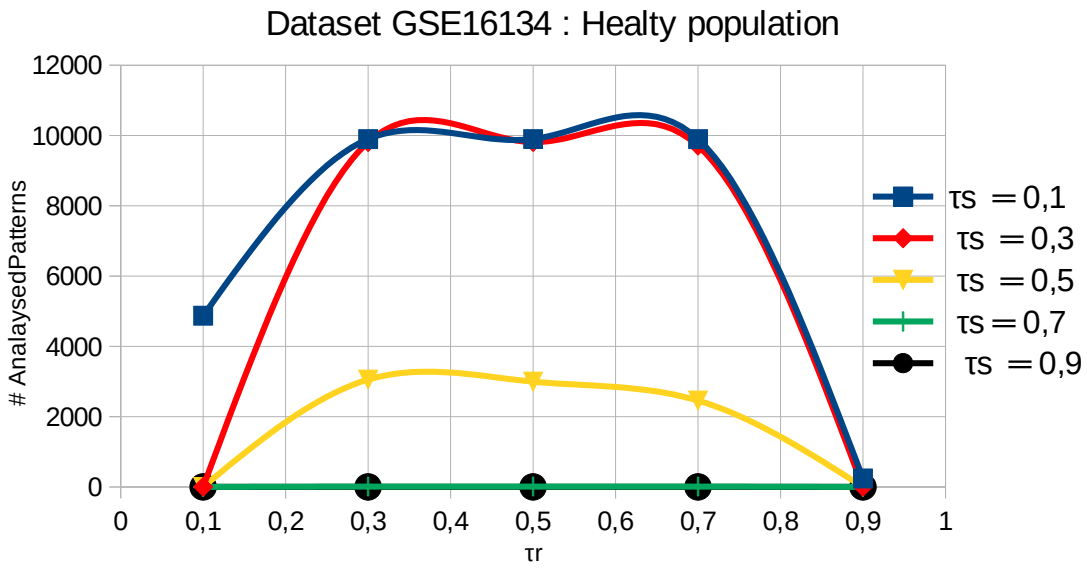
Anche in questo caso, dai file di output restituiti dopo l'invocazione dell'algoritmo di mining, si dovrà estrarre il valore corrispondente alla dicitura Visited Nodes, indicativo del numero di sottografi esaminati per ognuna delle due reti.

I risultati ottenuti verranno esplicati tracciando una coppia di grafici per ogni dataset analizzati e nello specifico:

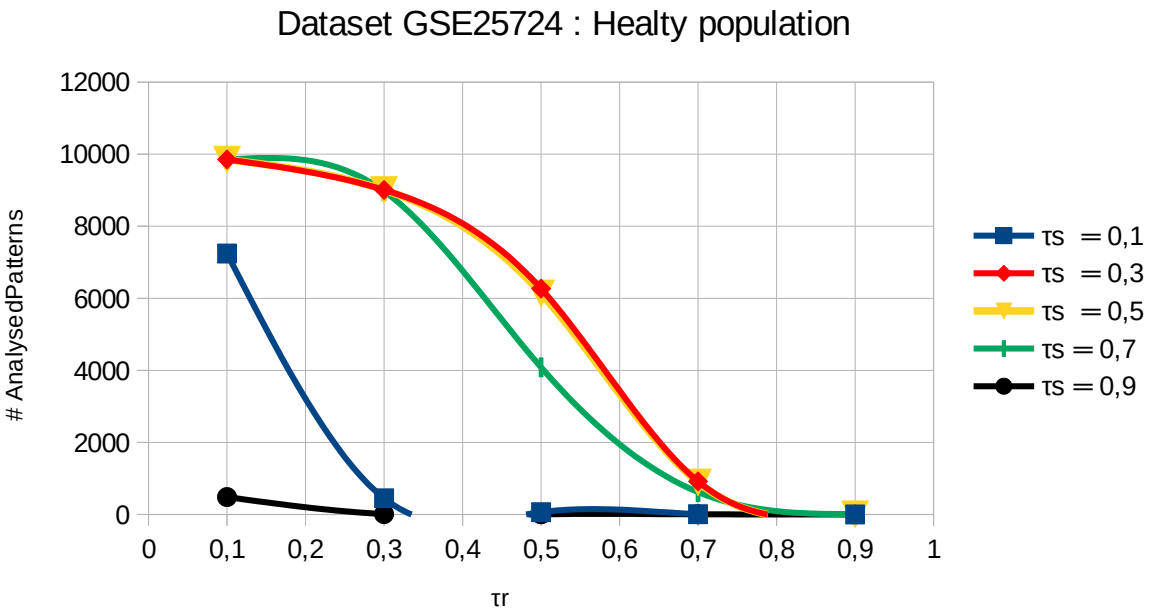
- un grafico che rappresenti l'andamento del numero di patterns analizzati, al variare delle due soglie, nel caso della rete associata agli individui sani, denominata Healty population;
- un grafico che descriva l'andamento della medesima funzione, nel caso, però, del grafo raffigurativo dei campioni estratti dai pazienti malati, definito come Unhealty population;

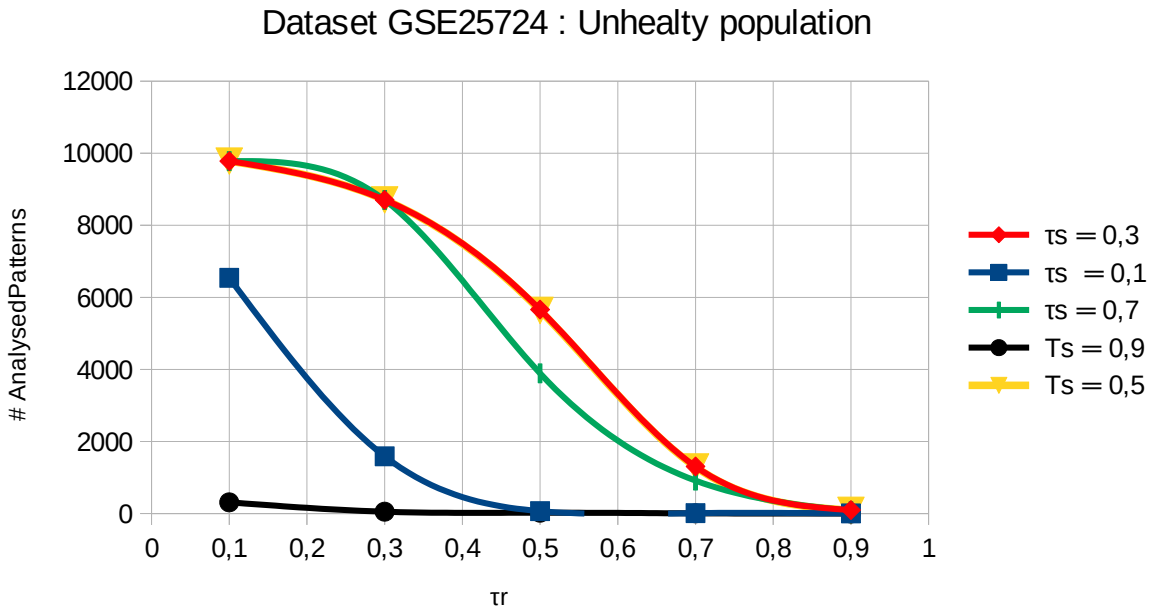
Si riportano i grafici corrispondenti, per quanto riguarda i vari datasets:

Dataset GSE16134

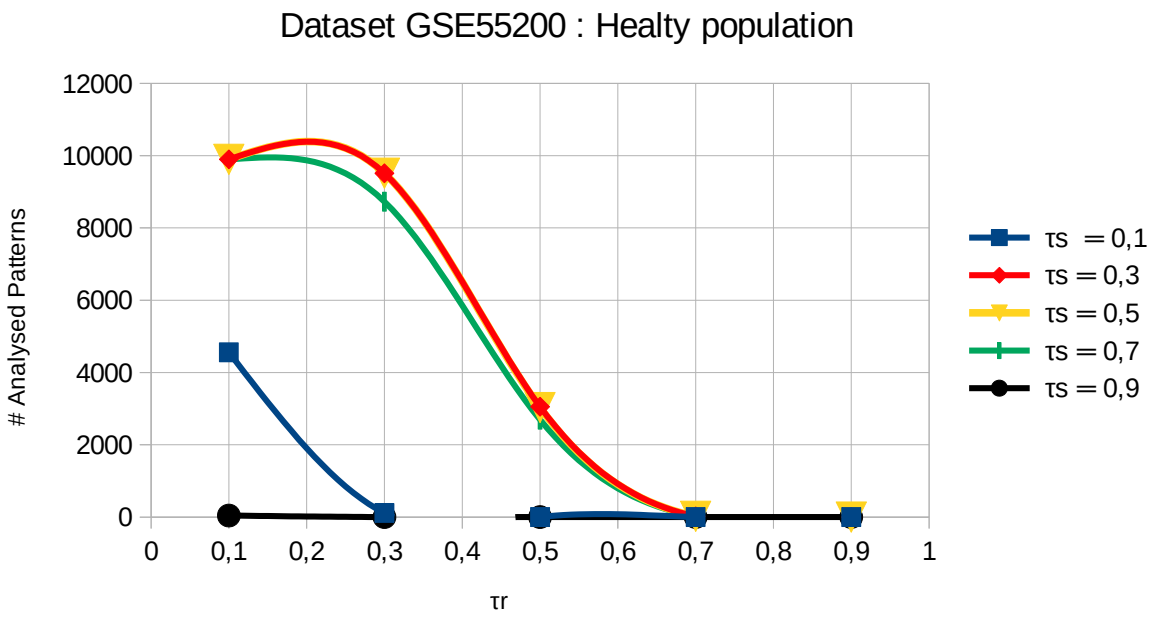


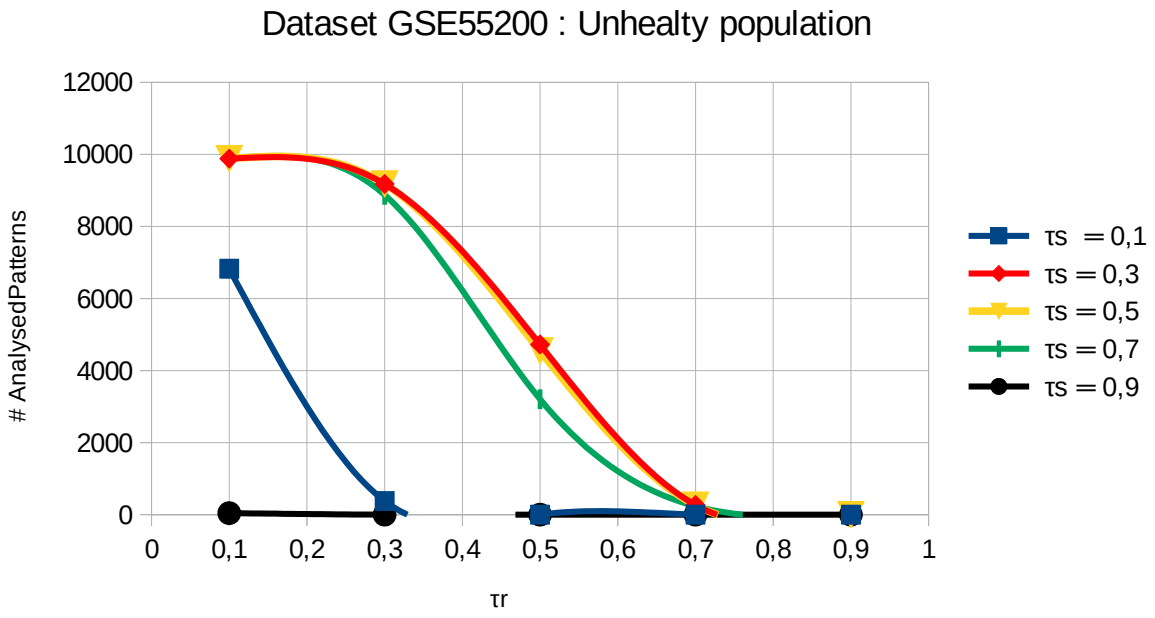
Dataset GSE25724



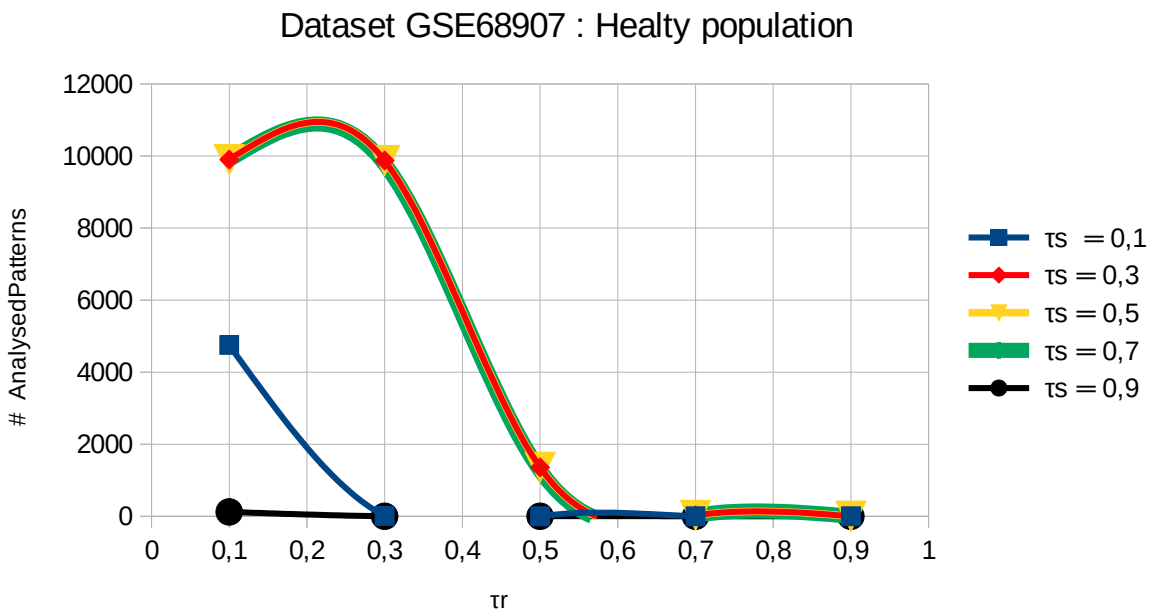


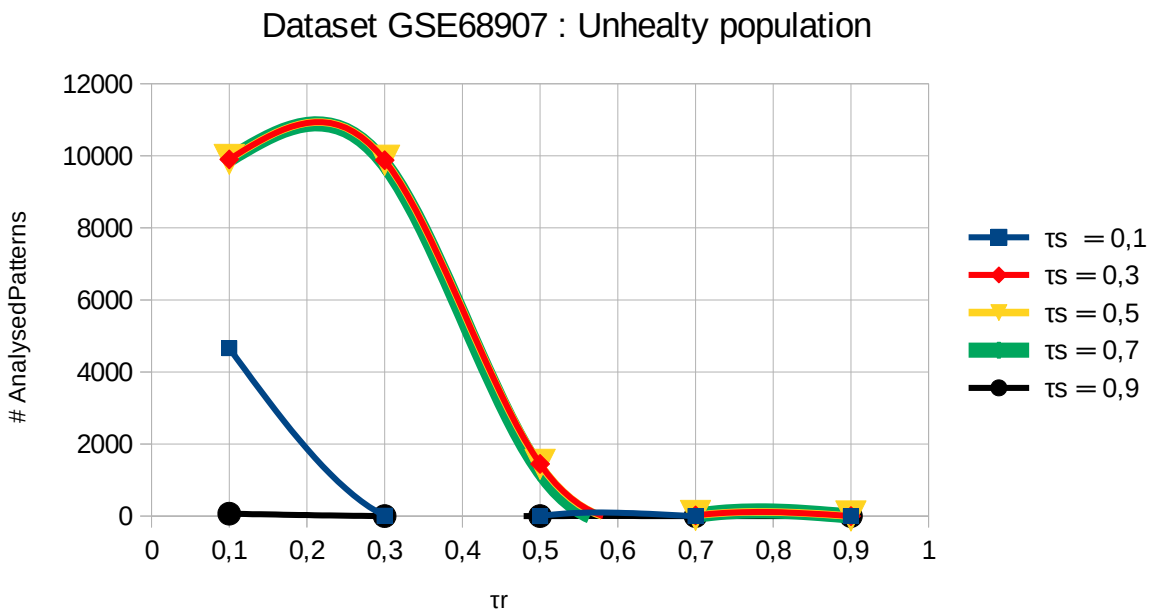
Dataset GSE55200





Dataset GSE68907


Dataset GSE68907 : Healty population
AnalysedPatterns
12000
10000
8000
6000
4000
2000
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
τr
τs = 0,1
τs = 0,3
τs = 0,5
τs = 0,7
τs = 0,9


Dataset GSE68907 : Unhealty population
AnalysedPatterns
12000
10000
8000
6000
4000
2000
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
τr
τs = 0,1
τs = 0,3
τs = 0,5
τs = 0,7
τs = 0,9

Quello che si vuole evidenziare, tramite i grafici riportati sopra, è la sensibilità che dimostra l'algoritmo alla variazione delle soglie "Strength Threshold" e "Relevance Threshold", in un intervallo sì fatto:

ThresholdVariation = { 0.1 ; 0.3 ; 0.5 ; 0.7 ; 0.9 }

eseguendo anche un confronto, tra il grafo dei campioni sani e quello dei malati, su come il numero di patterns trovati, nell'una e nell'altra rete ,differisca in base ai valori assunti dalle soglie.

Riferimenti

[1] D. J. Watts, S.H. Strogatz, “Collective dynamics of ‘small-world’”, in Nature.

[2] D. J. Watts, M.Newman, “Identity and Search in Social Networks”, in Science (2002);

[3] Erdős, Rényi, (1959); A. L. Barabási, (2002);

[4] V.Latore, “Reti small world: l’architettura di un sistema complesso”;

[5] D. Fell, A. Wagner, “The small world of metabolism”, in Nature Biotechnology (2000);

[6] D. Fell, A. Wagner, “The small world inside large metabolic networks”, (2001);

[7] R. Albert, A. L. Barabási, “Statistical mechanics of complex networks”, (2002);

[8] L. H. Hartwell, J. J. Hopfield, S. Leibler, A. W. Murray, “From molecular to modular cell biology”;

[9] E. de Silva, M. P.H Stumpf, “Complex networks and simple models in biology”, (2006);

[10] F.Fassetti, S. E. Rombo, C. Serrao, “Discriminative Pattern Discovery on Biological Networks”, (2017);

[11] R. Saito, H. Suzuki, Y. Hayashizaki, “Interaction generality, a measurement to assess the reliability of a protein–protein interaction”;

[12] A. L. Barabási, Z. N. Oltvai,“Network Biology: Understanding The Cell's Functional Organization”;

[13] Y. Kim, T. M. Przytycka, “Bridging the gap between genotype and phenotype via network approaches”;

[14] H. Carter , M. Hofree , T. Ideker, “ Genotype to phenotype via network

analysis";

[15] M. Vidal, M. E. Cusick, A. L. Barabási, "Interactome Networks and Human Disease";

[16] O. B. Poirion, X. Zhu, T. Ching, L. X. Garmire, "Using Single Nucleotide Variations in Single-Cell RNA-Seq to Identify Tumor Subpopulations and Genotype-phenotype Linkage";

[17] N. Atias, R. Sharan, "Comparative analysis of protein networks: hard problems, pratical solutions";

[18] A. J. Butte, I. S. Kohane, "Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks";

[19] J. A. Eisen, "Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis";

[20] S. F. Altschul, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs";

[21] Roy, Swarup; Bhattacharyya, Dhruba K; Kalita, Jugal K, "Reconstruction of gene co-expression network from microarray data using local expression patterns", (2014);

[22] M. Lajoie, D. Bertrand, N. El-Mabrouk,"Inferring the Evolutionary History of Gene Clusters from Phylogenetic and Gene Order Data", (2010);

[23] M. C. Oldham, S. Horvath, D. H. Geschwind, "Conservation and evolution of gene coexpression networks in human and chimpanzee brains";

[24] K. Aoki, Y. Ogata, D. Shibata, "Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology2;

[25] S. A. Cook, "The complexity of theorem-rpoving procedures", (1971);

[26] C. Lin, Y. Cho, W. Hwang, P. Pei, A. Zhang, "Knowledge Discovery in Bioinformatics: Techniques, Methods and Application", (2006);

[27] S. S. Shen-Orr1, R. Milo, S. Mangan, U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli";

[28] G. Ciriello, C. Guerra, "A review on models and algorithms for motif discovery in protein-protein interaction networks",(2008);

[29] N. Kashtan, S. Itzkovitz, R. Milo, U. Alon, “Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs”;

[30] T. I. Lee1, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hann, “Transcriptional Regulatory Networks in Saccharomyces cerevisiae”;

[31] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, “Network Motifs: Simple Building Blocks of Complex Networks”;

[32] P. G. Ferreira, P. J. Azevedo, “Evaluating Protein Motif Significance Measures: A Case Study on Prosite Patterns”;

[33] G. Ferreira, P. J. Azevedo, “Evaluating deterministic motif significance measures in protein databases”;

[34] V. Lacroix, C.G. Fernandes, Marie-France Sagot, “Motif Search in Graphs: Application to Metabolic Networks”;

[35] S. Mangan, U. Alo, “Structure and function of the feed-forward loop network motif”;

[36] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, Uri Alon, H. Margali, “Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction”;

[37] J. Ruan, W. Zhang, “Identifying network communities with a high resolution”;

[38] A. Enright, S. Van Dongen, “An efficient algorithm for large-scale detection of protein families”;

[39] S. Rombo, C. Pizzuti, “Experimental Evaluation of Topological-based Fitness Functions to Detect Complexes in PPI Networks”;

[40] P. Kaur, K. Si. Attwal, “Data Mining:Review “;

[41] C. C. Aggarwal, J. Han, “Frequent Pattern Mining”;

[42] J. Han, H. Cheng, D. Xin, X. Yan, “Frequent pattern mining: current status and future directions”;

[43] Park, Chen, Yu, “An effective hash-based algorithm for mining

association rules", (1995);
[44] A. Savasere, E. Omiecinski, S. Navathe, "An efficient algorithm for mining association rules in large databases", (1995);
[45] H. Toivonen, "Sampling large databases for association rules", (1996);
[46] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation", (2000);
[47] V. Porcu, "Introduzione al machine learning con R";
[48] J. Bailey, T. Manoukian, K. Ramamohanarao, "Fast algorithms for mining emerging patterns", (2002);
[49] J. Li, L. Wong, "Emerging patterns and gene expression data", (2001);
[50] K. Ramamohanarao, J. Bailey, H. Fan, "Efficient mining of contrast patterns and their applications to classification", (2005);
[51] A. Barabási, N. Gulbahce, J. Loscalzo, "Network Medicine: A Network-based Approach to Human Disease";
[52] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, A. Barabási, "The human disease network";
[53] J. C. Chen, M. J. Alvarez, F. Talos, "Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks", (2014);
[54] Z. Shao, Y. Hirayama, Y. Yamanishi, H. Saigo, "Mining discriminative patterns from graph data with multiple labels and its application to quantitative structure-activity relationship (QSAR) models", (2015);
[55] Z. Wang, Y. Zhao, G. Wang, Y. Li, X. Wang, "On extending extreme learning machine to non-redundant synergy pattern based graph classification", (2015);
[56] X. Yan, H. Cheng, J. Han, P. S. Yu, "Mining significant graph patterns by leap search", (2008);
[57] C. Prieto, A. Risueno, C. Fontanillo, J. De Las Rivas, "Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles";

[58] D. Anastassiou, “Computational analysis of the synergy among multiple interacting genes”;

[59] Gray, “Entropy and information theory”;