POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Biomedica

Tesi di Laurea Magistrale

ANALISI DI RETI DI CO-ESPRESSIONE GENICA COINVOLTE NELL'AUTISMO

Relatore
Prof. Alfredo Benso

Correlatori
Prof. Stefano Di Carlo
Prof.ssa Gabriella Olmo
Prof. Gianfranco Michele Maria Politano

Candidate
Margherita Isola s242436
Annalisa Letizia s244448

Anno Accademico 2018-2019

Sommario

Il disturbo dello spettro autistico incide sul comportamento, l'interazione sociale e la comunicazione da parte del soggetto; è un disturbo ad alta incidenza, che può pregiudicare gravemente il tenore di vita dei soggetti e rappresenta una grande incognita nel mondo della medicina. Non è stata, infatti, ancora identificata una causa specifica nello sviluppo di questo disturbo e le diagnosi odierne si basano su valutazioni comportamentali che possono essere estremamente influenzate dalla soggettività sia dell'analista che del soggetto. Esso coinvolge le funzioni articolate del cervello in fase di sviluppo, le quali potrebbero essere influenzate da un enorme pool di fattori. Ancora non è stato trovato un rapporto di casualità tra autismo e mutazioni, sia singole che combinate, sebbene siano diverse le mutazioni associate all'autismo, le quali non riescono a garantire ripetibilità e specificità.

Il lavoro di tesi si prefigge, dunque, lo scopo di studiare reti di co-espressione e anti-correlazione genica e studiare l'impatto della variazione dell'espressione di un determinato pool genico potenzialmente coinvolto nella diagnosi di autismo. Il presente lavoro di tesi è articolato come segue:

I Introduzione al problema, focalizzato sull'autismo e sullo stato dell'arte della diagnosi e delle tecniche di neuroimaging e genetica; al mondo della bioinformatica, della systems biology e della teoria dei grafi, rappresentanti i metodi di analisi del problema.

II Implementazione di algoritmi per la semplificazione e per la visualizzazione più comprensibile di grafi da parte di una classe di utenti con una specializzazione in biologia, o sprovvisti di conoscenze profonde in ambito informatico. In particolare, si sono isolate diverse caratteristiche dei dati, come: le relazioni tra le varie entità sia a livello di signalling che a livello biochimico, il numero di pubblicazioni, la presenza di complessi proteici e l'individuazione di entità con grande interesse biologico. Tutte queste divisioni vengono rappresentate attraverso un grafico che permette una visione più comprensibile dei dati a disposizione.

III Implementazione di simulatori che permettono la propagazione dell'espressione di geni. Più specificatamente, si è forzata l'attivazione e l'inibizione di un nodo cardine e si è propagato il suo effetto verso le altre entità della rete. Sono stati progettati due simulatori:

- Il primo simulatore rappresenta un prototipo e prevede una propagazione semplice dell'espressione del nodo sorgente fino ad un nodo identificato come target, seguita da un confronto degli stati di tutte le entità con un documento considerato Golden Standard, contenente tutte le espressioni dei nodi, permettendo dunque di discriminare le varie relazioni come corrette o sbagliate. Il presente simulatore è stato considerato prototipo, non essendo stato possibile ottenere il Golden Standard con il quale confrontare i risultati ottenuti dal forzamento delle espressioni.
- Il secondo simulatore invece risulta essere più realistico, non necessitando di dati aggiuntivi che confermano o meno l'espressione di ogni singola entità. Esso forza l'espressione di un determinato pool di nodi e va ad analizzare, singolarmente, i percorsi rappresentanti cause ed effetti dovuti al forzamento di tale espressione, riuscendo poi a restituire un grado di coerenza di ogni arco, rapportato al percorso di estrazione, rappresentato da percentuali di correttezza, incertezza e dubbiosità, nonché uno stato predetto, ottenuto analizzando le percentuali e confrontandole con delle soglie. Questo simulatore sarà dunque identificato come definitivo.

IV Applicazione del simulatore ad un pool di geni coinvolti nell'autismo, interpretazione ed analisi dei risultati finali ottenuti dalle simulazioni.

V Conclusioni e sviluppi futuri.

Il grafo iniziale è stato estratto dal database TheRingDB. Tutti i codici sorgente sono state sviluppati mediante il linguaggio Python (in versione 2.7 e 3.8), sfruttando le librerie Networkx e Pandas.

Abstract

The autistic spectrum disorder affects behavior, social interaction and communication by the subject; it is a high-incidence disorder, which can seriously affect the standard of living of the subjects and represents a great unknown in the world of medicine. In fact, a specific cause in the development of this disorder has not yet been identified and today's diagnoses are based on behavioural evaluations that can be extremely influenced by the subjectivity of both the analyst and the subject. It involves the articulated functions of the developing brain, which could be affected by a huge pool of factors. A causal relationship between autism and mutations, both single and combined, has not yet been found, although there are several mutations associated with autism, which fail to guarantee repeatability and specificity.

The aim of thesis is, therefore, to study networks of co-expression and anti-correlation gene and study the impact of the variation of the expression of a certain gene pool potentially involved in the diagnosis of autism. This thesis work is structured as follows:

I Introduction to the problem, focusing on autism and the state of the art of diagnosis and techniques of neuroimaging and genetics, and the world of bioinformatics, systems biology and graph theory, the methods of analysis of the problem.

II Implementation of algorithms for simplification and more understandable visualization of graphs by a class of users with a specialization in biology, or without deep knowledge in the field of computer science. In particular, several characteristics of the data have been isolated, such as: the relationships between the various entities at the level of signalling as well as at the biochemical level, the number of publications, the presence of protein complexes and the identification of entities with great biological interest. All these divisions are represented through a graph that allows a more comprehensible view of the available data.

III Implementation of simulators allowing the propagation of gene expression. More specifically, it forced the activation and inhibition of a pivotal node and propagated its effect to the other entities of the network. Two simulators have been designed:

- The first simulator represents a prototype and provides simple propagation of the expression of the source node to a target node, followed by a comparison of the states of all the entities with a document considered Golden Standard, containing all the expressions of the nodes, thus allowing to discriminate the various relationships as correct or wrong. This simulator was considered a prototype, as it was not possible to obtain the Golden Standard, with which to compare the results obtained by forcing expressions.
- The second simulator, on the other hand, turns out to be more realistic, not requiring additional data that confirm or not the expression of each entity. It forces the expression of a certain pool of nodes and goes to analyze, individually, the paths representing causes and effects due to the forcing of such expression, succeeding then to restore a degree of coherence of each arc, in relation to the extraction path, represented by percentages of correctness, uncertainty and doubtfulness, as well as a predetermined state, obtained by analysing percentages and comparing them with thresholds. This simulator will therefore be identified as definitive.

IV Application of the simulator to a pool of genes involved in autism, interpretation and analysis of the final results obtained from simulations.

V Conclusions and future developments.

The initial graph was pulled from the TheRingDB database. All source codes were developed using the Python language (version 2.7 and 3.8), using the Networkx and pandas libraries.

Indice

Elenco delle figure

Elenco delle tabelle

Parte I

Introduzione

Capitolo 1

Lo spettro autistico

Dati e diagnosi

Il *disturbo dello spettro autistico*, come riportato dal *NIMH (National Institute of Mental Health)* [15], è identificato come un disturbo dello sviluppo, data la precocità dell'apparizione dei suoi sintomi (pur non escludendo la possibilità di una diagnosi tardiva), che ha forte impatto sulla carattere, sul comportamento e sulle abilità comunicative del soggetto. Esso è definito come uno spettro poiché internamente alla sua macro-aeree include altre micro-tipologie, dovute alle sue diverse variazioni e alla gravità dei sintomi.
Le persone affette da disturbo dello spettro autistico mostrano in particolar modo difficoltà nella comunicazione e nell'interazione sociale, atteggiamenti assorbenti e ristretti, nonché comportamenti ripetitivi e stereotipie, che hanno forte impatto nella gestione della vita quotidiana.

Sebbene non ci siano dati certi sulla causa dell'autismo, la cooperazione tra geni e le influenze esterne sembrano poter avere delle conseguenze che potrebbero portare alla diagnosi di autismo, creando così fattori di rischio, quali l'età dei genitori, l'appartenenza ad una famiglia con altri casi di autismo, co-morbilità con diverse e specifiche condizioni e patologie genetiche e particolari fattori verificatisi al momento della nascita. Nonostante ciò, non essendoci né una diagnosi sufficientemente accurata, come si vedrà successivamente, né la determinazione di una causa certa, i trattamenti medici risultano essere poco efficienti in termine di targeting attivo e specifico, dividendosi principalmente in terapie farmacologiche, necessarie per l'attenuazione e il trattamento di determinati sintomi, e terapie comportamentali, utili a migliorare la qualità di vita e il self-empowerment del soggetto.

L'incidenza dell'autismo sulla popolazione è incrementata rapidamente nel corso degli anni, come è possibile vedere nelle statistiche riportate in [9], passando da un ratio iniziale di 1 su 150 ad uno di 1 su 59.
Questi valori sono ricavati da alcuni studi effettuati biennalmente dal *CDC (Centers for Disease Control and Prevention)*, usando un sistema, chiamato *Autism and Developmental Disabilities Monitoring Network (ADDM Network)*, in grado di fornire delle stime riguardanti la diffusione di disturbi dello spettro autistico (DSA) tra bambini di 8 anni i cui genitori/tutori vivono negli 11 Stati degli USA (Arizona, Arkansas, Colorado, Georgia, Maryland, Minnesota, Missouri, New Jersey, North Carolina, Tennessee, Wisconsin) in cui viene condotto questo studio pilota.

L'esecuzione del monitoraggio è effettuato in 2 step distinti:

- vengono raccolti dati provenienti da diverse risorse e report, eseguiti da providers professionali istruiti e supervisionati, al fine di certificarne la formazione iniziale, identificare esigenze attuali e garantire l'adesione alle metodologie richieste;
- i dati raccolti precedentemente vengono studiati e processati al fine di identificare casi di DSA, congruenti ai criteri diagnostici forniti dal DSM.

L'ultimo studio pubblicato [3] risalente al 2014 prende come riferimento il DSM-IV-TR, nonostante nel 2013 fosse stata pubblicata, da parte dell'*American Psychiatric Association*, la nuova edizione del DSM, il DSM-V; per questo motivo, molti campioni analizzati nello studio sono stati sottoposti ad ulteriori accertamenti in caso di validazione e di congruenza con i criteri diagnostici del DSM-V, permettendo ai bambini la classificazione secondo uno o entrambi i criteri, quali:

1. comportamenti congruenti ai criteri del DSM-V;
2. diagnosi di DSA secondo il DSM-V o DSM-IV-TR.

La prevalenza globale di DSA è pari al 16.8 per 1000 bambini di 8 anni, variando in un range compreso tra 13.1 e 29.3.
Sono state trovate delle influenze dovute al sesso e all'etnia di appartenenza: infatti, i maschi risultano essere 4 volte più sensibili alla diagnosi rispetto alle donne; mentre i bambini bianchi non ispanici risultano essere più sensibili rispetto ai bambini neri non ispanici, ed entrambi risultavano essere più inclini rispetto agli ispanici. Anche il QI medio ha subito variazioni nei diversi siti, dipendente anch'esso dal sesso e dalla razza. L'età alla quale è stata effettuata la prima diagnosi tuttaviam è risultata mediamente pari al 52esimo mese di vita, non mostrando grandi variazioni in base al sesso e l'etnia.
I dati risultavano affini passando dai criteri diagnostici del DSM-IV-TR a quelli del DSM-V, con un surplus di casi appartenenti al DSM-IV-TR di meno del 5% ed una sovrapposizione dell'86%.
La prevalenza globale è indubbiamente aumentata rispetto a quanto riportato negli studi precedenti [9], ma bisogna segnalare che lo screening [1] non può essere totalmente rappresentativo di quanto accade ai bambini appartenenti agli USA, includendo solo una porzione ridotta del paese. Ciò nonostante, una prevalenza così alta richiede un'aumento di servizi e il progresso della ricerca riguardante il DSA e le probabili cause.

La diagnosi, come già detto in precedenza, viene effettuata tutt'oggi semplicemente basandosi sulla valutazione e l'osservazione del soggetto, due misure che risultano essere poco oggettive, portando ad una diagnosi poco attendibile. Inoltre, la valutazione dei soggetti, essendo allineata al DSM, è stata ed è notevolmente mutevole, come riportato da Temple Grandin nel suo libro [46].

Per comprendere appieno la complessità di valutazione è opportuno ripercorrere nel tempo l'evoluzione nella diagnosi.

1943, Leo Kanner: il medico della Johns Hopkins University ha condotto la prima diagnosi di autismo in uno studio relativo ad 11 bambini che condividevano sintomi che sarebbero poi diventati distintivi dell'autismo. In questo studio, Kanner prepondeva per una spiegazione biologica alla base dell'autismo, essendo diagnosticabile e presente già a un'età precoce e osservavando analogie tra i comportamenti dei genitori, i quali condividevano con i figli il pool genico.

1949, Leo Kanner: il focus di Kanner si spostò sulla psicologia, subendo l'influenza dell'affermarsi del pensiero psicoanalitico. Nel suo lavoro, concentrò la maggiorparte dei suoi sforzi nell'analisi dei comportamenti genitoriali, attribuendo la causa dell'autismo a comportamenti inopportuni messi in atto dai genitori, a scarse cure e alla distanza emotiva dei genitori stessi. Come riportato da Temple Grandin [46], sembra evidente che Kanner avesse scambiato cause ed effetti, non domandandosi se un eventuale atteggiamento di distanza da parte dei genitori, nei confronti del bambino, non fosse una risposta consequenziale all'incapacità di comunicare e di comprendere dello stesso e non viceversa. Anche Bettelheim, direttore della University of Chicago's Orthogenic School fu influenzato dal pensiero di Kanner e parlò di predisposizione del bambino nel manifestare sintomi autistici, non predeterminazione, come se la sindrome fosse latente e venisse risvegliata da mancanze genitoriali.

1952, DSM-I: come riportato da Steve Silberman [42], all'interno del DSM-I, l'autismo fa la sua prima comparsa come una *"schizofrenia, di tipo infantile"*, senza dare una definizione precisa di come essa si manifestasse.

1968, DSM-II: la descrizione della così chiamata *"schizofrenia, di tipo infantile"* compare solo successivamente, nel DSM-II, subendo però l'influenza "negativa" del pensiero di Bettelheim, indicando il "comportamento autistico, atipico e ritirato" come prova della mancanza di uno sviluppo separato rispetto a quello della madre.

1980, DSM-III: Il pensiero psichiatrico subì una netta svolta, modificando il proprio focus dalle cause agli effetti, in modo tale da poter classificare i sintomi in maniera più rigorosa, reinventando la psichiatria come una scienza medica esatta, interfaccia dell'industria farmaceutica. Spitzer basò così il DSM-III su una quantità quanto più grande di studi empirici, formando 25 commissioni, i cui membri furono definiti come "Data-Oriented People" (DOP), in grado di sviluppare una descrizione dettagliata dei diversi disturbi. L'autismo infantile, preannunciato da Kanner, fu così racchiuso in una categoria più ampia, nota come "Disturbi pervasivi dello sviluppo" (PDD), con l'avvento del DSM-III. Per ricevere una diagnosi di autismo infantile era necessario che il paziente soddisfacesse almeno 6 criteri, tra i quali l'assenza di sintomi rimandanti alla schizofrenia e l'esordio precedente ai 30 mesi di vita. Al fine di tutelare i bambini che avevano sofferto una successiva perdita di capacità, fu introdotto il "Disturbo pervasivo dello sviluppo a esordio infantile" (COPDD).

1987, DSM-III-R: al DSM-III venne però attribuita la colpa di includere, per l'autismo, criteri difficili da applicare nella pratica, richiedendo una revisione del lavoro, chiamata DSM-III-R. Il nome venne così cambiato, da *"Autismo infantile"* a *"Disturbo autistico"*, pur aumentando il numero dei criteri diagnostici, da 6 a 16 criteri, distinti in 3 categorie e specificando quanti criteri dovesse soddisfare un soggetto per ogni categoria, al fine di ottenere la diagnosi, generando così, come riportato da Temple Grandin [46], un incremento vertiginoso di diagnosi di autismo. La categoria dei PDD venne inoltre ampliata dalla categoria dei PDD-NOS (Disturbi pervasivi dello sviluppo non altrimenti specificati), alla quale appartenevano i casi in cui i sintomi autistici erano lievi o insufficienti ad ottenere la diagnosi.

1994, DSM-IV: in questa versione compare una nuova diagnosi, nota come *"Disturbo di Asperger"*, ottenuta dagli studi del 1943 e 1944 del pediatra austriaco Hans Asperger, che identificava una nuova categoria di bambini presentanti determinati comportamenti. Furono proprio queste aggiunte, unitamente alla presenza del *"Disturbo di Rett"* e *"Disturbo disintegrativo dell'infanzia"*, a permettere all'autismo di presentarsi come uno spettro.

2000, DSM-IV-TR: correzione dei criteri diagnostici un alcune patologie.
Nel 2012 il DSM-IV-TR è stato rinnovato con il DSM-V, dove Il *"Disturbo di Asperger"* si trasformò in *"Autismo ad alto funzionamento"*, e con questa nuova versione del DSM si parla indistintamente di PDD e DSA.

L'accessibilità da parte della comunità di una nuova diagnosi, gli standard meno rigorosi l'ampiezza dello spettro e l'inclusione di Asperger, PDD-NOS e DSA, l'accresciuta sensibilità e consapevolezza del problema portò, come riferito da Temple Grandin [46], ad una vera e propria "epidemia di autismo".
Grazie ai progressi tecnologici, in termini di neuroimaging e genetica, la scienza potrebbe permetterci nel tempo di rispondere a due domande irrisolte a causa di una diagnosi ancora troppo soggettiva, mutevole e non basata su dati biologici e certi: Cosa fa l'autismo? Come lo fa?

Applicazioni di neuroimaging e genetica, mirate ad una diagnosi più accurata

Le tecniche di neuroimaging, in particolare la Risonanza Magnetica (MRI), permettono di illustrare l'aspetto di un cervello e le azioni svolte, pur non permettendo una distinzione tra cause ed effetto.
La MRI sfrutta un magnete e particolari onde provenienti a frequenze radio specifiche in modo tale da consentire la rilevazione degli atomi di idrogeno del nostro corpo. Mentre la risonanza magnetica strutturale (sMRI) permette di illustrare le strutture anatomiche, la risonanza magnetica funzionale (fMRI) permette di illustrarne il funzionamento in risposta a determinati stimoli e azioni.
Prima del neuroimaging, per la spiegazione anatomica i ricercatori dovevano ricorrere all'esame autoptico, mentre la risposta funzionale era quasi impossibile da ottenere.

Temple Grandin, la professoressa associata della Colorado State University, si sottopose a diverse scansioni del suo cervello [15], ottenendo informazioni in grado di spiegare parzialmente alcuni suoi comportamenti in quanto autistica.
Coordinazione motoria:
Il suo cervelletto, responsabile della coordinazione motoria, risultava essere inferiore del 20% rispetto ad un soggetto di controllo, spiegando i problemi motori frequenti in persone con DSA.
Abilità visive:
Successivamente, fu in grado di scoprire tramite fMRI che l'attivazione della sua corteccia visiva ventrale rispondeva in modo analogo guardando disegni di oggetti rispetto ad un soggetto di controllo, mentre rispondeva in modo minore a volti umani; inoltre, tramite imaging con tensore di diffusione (DTI), fu in grado di scoprire un'interconnessione nel suo cervello tra il fascicolo fronto-occipitale inferiore (IFOF) e il fascicolo longitudinale inferiore (ILF), spiegando probabilmente le grandi capacità in termini di memoria visiva da parte di soggetti con DSA.
Abilità matematiche:
Successive scansioni analizzarono l'asimmetria tra i 2 ventricoli del suo cervello, pari al 55%, rispetto al 15% di un soggetto di controllo, provocando una estensione del ventricolo sinistro all'interno della corteccia parietale, responsabile della memoria a breve termine e delle abilità matematiche, con la quale hanno difficoltà molti soggetti con DSA. Probabilmente la differenza si potrebbe spiegare con un tentativo di compensazione del cervello, di un danno cerebrale nelle fase primordiali dello sviluppo.
Ansia:
Il volume intracranico, le dimensioni del suo cervello, la quantità di materia bianca risultavano maggiori del 15% rispetto a un soggetto neurotipico, simboleggiando con alta probabilità il tentativo di compensazione del danno cerebrale operato dal cervello. Anche le amigdale, fondamentali per l'elaborazione di alcuni stati d'animo, risultavano più grandi nel normale, spiegando i problemi d'ansia presenti in pazienti con DSA.

Abilità mnemoniche:
Lo spessore delle corteccie entorinali, in ultima battuta, risultava essere maggiore del 12% rispetto ai soggetti neurotipici, giustificandone ulteriormente le eccezionali capacità mnemoniche. Molti di questi fattori risultano essere condivisi da altre persone con DSA, confermando la presenza dell'autismo nel cervello.

La difficoltà nella standardizzazione risiedono in 3 fattori:

- Omogeneità delle strutture cerebrali: Molte anomalie anatomiche, riscontrate in cervelli come quello di Temple Grandin, non risultano essere presenti in tutti i pazienti con DSA nei quali le differenze tendono a posizionarsi nell'ambito della normalità, classificando anatomicamente i pazienti con DSA e i loro cervelli come normali. Nonostante ciò, vi sono alcune regolarità, alcuni pattern ampiamente condivisi tra i pazienti autistici.

- Eterogeneità delle cause: Pur ritrovando correlazioni tra comportamenti ed anomalie, non è detto che la stessa anomalia debba essere presente in alte persone manifestanti lo stesso atteggiamento.

- Eterogeneità dei comportamenti: Un'anomalia presente nel cervello, può non avere lo stesso effetto applicata ad un altro cervello.

Ciò nonostante, la neuroanatomia e la combinazione di dati potrebbero diventare un ottimo strumento diagnostico, permettendo una predizione delle difficoltà e una impostazione ottimale di terapie mirate.
L'obiettivo risulta essere l'identificazione non del cervello autistico, ma di elementi tali da fungere da marker biologici.
Ulteriore obiettivo è la scansione del cervello in maniera più dettagliata, come nel caso dell'*High-Definition Fiber Tracking* (HDFT), con la quale Temple Grandin [46] ha potuto sapere di avere un tratto visivo fuori dal comune, avvalorando ancora una volta le sue capacità mnemoniche, e una connessione tra "visto" e "parlato" pari all'1% rispetto ad un soggetto di controllo, giustificandone le scarse capacità in merito alla produzione del linguaggio.
Tutto ciò non ci permette di dare una spiegazione, una causa certa nascosta dietro l'autismo, bensì una indicazione anatomica degli effetti.
Al fine di ritrovare la causa, ci si rivolge alla genetica.
I primi risultati utili per la ricerca genetica nel campo dell'autismo sono attribuibili all'iniziativa federale *ENCODE (Encyclopedia of DNA Elements)*, per merito della quale è stato possibile iniziare a spiegare il ruolo del genoma, il cui aspetto era già stato precedentemente indagato dallo *Human Genome Project*.
Considerando la conformazione a doppia elica strettamente avvolta, propria del DNA, è stato possibile considerare l'effetto, in termini di regolazione, che la vicinanza di due frammenti distinti di DNA potevano avere tra di loro: infatti, un frammento distante migliaia di coppie di basi da un altro, in termini di sequenza sulla singola elica, poteva in realtà ritrovarsi vicino a questo nel contesto dello strettissimo avvolgimento nel quale si trova il DNA, azionando eventualmente "interruttori" presenti su di esso, ponendo i geni e gli elementi regolatori nel contesto tridimensionale in cui sono naturalmente. Sebbene la genetica sia ancora una scienza "primordiale", la sola conoscenza della possibilità che essa svolga un ruolo nell'autismo risulta essere un enorme passo avanti.

Gli studi genetici riferiti al campo del disturbo dello spettro autistico sono stati numerosi negli anni:

- **1977**:
  Come riferito da Temple Grandin [46] risale al 1977 il primo studio genetico focalizzato sull'autismo nei gemelli, il quale permise di scoprire, sebbene con un campione limitato, che il tasso di concordanza tra coppie di gemelli identici era del 36%, mentre per i gemelli fraterni era nullo.
  Oggigiorno, il tasso di concordanza sarebbe dell'86% e del 10% rispettivamente per i gemelli identici e fraterni.

- **1995**:
  In uno studio successivo del 1995, venne utilizzato un campione di misura doppia rispetto a quello di 18 anni prima, ottenendo risultati paragonabili e avvalorando la tesi dell'origine genetica dell'autismo, dal momento in cui i gemelli condividono lo stesso DNA. Nonostante ciò, il tasso di concordanza non risulta pari al 100%, poiché sebbene il genotipo sia identico nei gemelli, i geni potrebbero funzionare in modo diverso; il genotipo potrebbe anche non essere identico, per via di mutazioni spontanee. Queste differenze genetiche formano infatti il fenotipo, responsabile dell'aspetto, intelletto, carattere del soggetto, nel quale i due gemelli potrebbero distinguersi.

- **2001**:
  Successivamente alla mappatura del genoma nel 2001 da parte dell'*HGP* e di *Celera Genomics*, diverse istituzioni appartenenti a 19 paesi si sono consorziate nell'*Autism Genome Project (AGP)*, al fine di ricercare le basi genetiche dell'autismo, utilizzando un database di 1400 famiglie e il gene chip, che ha permesso di osservare varianti di DNA contemporaneamente.

- **2007**:
  Nel 2007, l'AGP pubblicò un articolo nel quale era segnalata una mutazione del gene codificante la neurexina, linkata alla neurolingina per controllare la connessione sinaptica tra le cellule, convalidando la ricerca precedente che indicava che sono associate al rischio di autismo mutazioni del gene codificante la proteina SHANK3, la quale interagisce con la neurolingina. Questi studi si sono basate sulla ricerca di copy number variations (CNV), le quali possono essere ereditarie o generarsi spontaneamente prima della fecondazione o subito dopo (de novo). Un altro studio del 2007, basato su 264 famiglie indicava che il 10% di bambini autistici con fratelli senza DSA avevano delle CNV de novo, che si verificavano nell'1% dei soggetti neurotipici. Successivamente furono associate all'autismo centinaia di CNV, che però non avevano alta ripetibilità, portando elementi che non si ripetevano più frequentemente dell'1% dei casi. L'eterogeneità dei tratti autistici complica le cose, sia in termini di differenze tra elementi autistici, sia in termini di condivisione di tratti con altri sindromi, richiedendo l'identificazione di caratteristiche principali e primarie per la condizioni autistica. Vennero così ricercati pattern all'interno delle mutazioni, trovando risultati interessanti riguardanti l'appartenenza di geni a categorie con effetto sulla proliferazione cellulare e la neurotrasmissione delle cellule cerebrali.

– **2012**:
In uno studio del 2012, tre gruppi di ricercatori crearono un pool di soggetti autistici con parenti prossimi senza comportamenti autistici e ricercarono le CNV de novo presenti in almeno 2 soggetti autistici e non presenti in alcun soggetto non autistico. Studi analoghi permisero la validazione di alcune CNV, ritrovate durante gli studi diversi e venne scoperto che esse probabilmente erano originate dal padre, rafforzando la cosa con uno studio sulla correlazione tra l'età del padre e il tasso di mutazioni de novo. Purtroppo, non è chiaro se la singola mutazione permette la creazione del tratto autistico o se esso è il risultato di una combinazione di mutazioni. La seconda considerazione fu rafforzata da molti studi, che scoprirono correlazioni tra CNV de novo rilevanti ed altre mutazioni frequentemente ripetute.

Anche i fattori ambientali assumono importanza nello studio genetico, interagendo con i geni causando alterazioni acquisite o somatiche. Diversi studi si focalizzarono sugli effetti delle alterazioni della gestazione, di fattori ambientali e farmacologici nell'ambito delle mutazioni geniche correlate all'autismo, trovando alcuni dati di validazione, come nel caso dell'assunzione dell'*acido valproico* durante la gestazione, riguardo il quale è possibile leggere nel sito SFARI che la correlazione tra assunzione da parte della mamma di acido valproico e incidenza di autismo nel figlio oscilla in un range del 6-9%.
Studi analoghi si verificarono nel caso di altri farmaci, come antidepressivi. Nonostante ciò, bisogna prestare attenzione a definire un rapporto di correlazione e confonderlo con una causazione.

L'esempio di predisposizione genetica maggiore è fornito, come riportato da Temple Gradin [46], dalla variante *7R del gene DRD4*, la quale rende il cervello degli individui meno sensibile alla dopamina, predisponendoli a disturbi della condotta e dell'attenzione.
Uno studio del 2010, riporta diverse associazioni tra bambini autistici con variante 7R e i loro genitori, identificando diverse combinazioni tra i tratti autistici e i genitori con la variante presente. Inizialmente era diffusa l'idea che i bambini con la variante 7R del gene fossero vulnerabili alle influenze negative dell'ambiente, semplicemente poiché si erano concentrate su di esse; successivamente si scoprì la neutralità del gene, che l'ambiente rendeva positivo o negativo, indicando, però, non una predisposizione all'autismo, come quella riportata da Kanner e Bettelheim, ma una predeterminazione.

Gli sviluppi della ricerca potrebbero rispondere a quesiti in grado di formulare una diagnosi ottimale, odiernamente basata solo sull'osservazione di comportamenti e, per questo, fallace e soggettiva.
É in questo campo d'azione, soprattutto nella genetica e, dunque, nella bioinformatica e systems biology, che si posiziona il nostro lavoro, con lo scopo di ritrovare peculiarità tipiche dell'autismo, analizzando network di co-espressione dei geni, al fine di segnalare path di interesse ai medici e ai biologi, che li approfondiranno secondariamente.

Capitolo 2

L'informatica a sostegno della biologia

I fattori genetici sono i maggiori contribuenti nello sviluppo della sindrome autistica da parte del bambino; ed è per questo che i biologi spingono sempre di più la loro ricerca verso l'individuazione dei geni partecipanti, in modo tale da ottenere in un futuro la diagnosi attraverso il solo studio dei geni del feto.
Questa possibilità oggigiorno ci sembra possibile e anche molto vicina alla realizzazione, ma per gli scienziati di cinquanta o sessanta anni fa tutto ciò sarebbe stata un'utopia.
In questi pochi anni è avvenuta un'evoluzione nel campo medico-biologico di una grandezza inimmaginabile, inoltre con l'ampliamento delle conoscenze e le scoperte rivoluzionarie su tutti i campi scientifici le variabili da considerare per poter anche solo osservare un problema sono diventate di un numero spropositato.
Pensiamo ad esempio a come prima per una diagnosi medica ci si basasse solo sulle sensazioni del paziente, mentre ora un dottore deve unire anche gli esami chimici del sangue e talvolta anche immagini.
Più si va avanti con le tecnologie e scoperte, più bisogna ampliare il proprio campo di osservazione ma anche unire diversi aspetti della scienza, come ad esempio la biologia e l'informatica. Unendo queste due grandi discipline, che ora coesistono ma che al tempo della nascita di una era inimmaginabile la possibile esistenza dell'altra, si ottiene la *bioinformatica*.
La bioinformatica negli anni ha subito numerose trasformazioni ed è arrivata ad intersecarsi in diversi ambiti, diventando talvolta anche strumento indispensabile per la comprensione dei dati e la risoluzione dei problemi.
Per capire meglio cos'è e che impatto ha avuto nel mondo scientifico è opportuno ripercorrere cronologicamente la sua evoluzione [20].

L'evoluzione dei dati biologici, dal 1866 al 2019

1866: la prima idea di genetica

Gregor Mendel è stato il fondatore della moderna genetica attraverso l'osservazione della famosissima pianta di *piselli odorosi*, vedi Figura 2.1.
Nasce così la prima idea di ereditarietà, di fenotipi, di caratteri dominanti e recessivi, in parole moderne: di genetica, [21]. Possiamo dire che la sue tabelle illustranti piselli e fiori è la primordiale rappresentazione di un database genomico, anche se a quel tempo Mendel non aveva minimamente idea di cosa fosse, infatti a quel tempo si parlava di caratteri e ancora non si sapeva che le informazioni osservate erano tutte trascritte nella doppia elica di acido desossiribonucleico presente in ogni cellula degli organismi viventi.

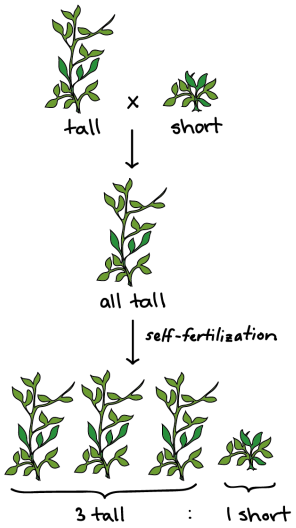


Figura 2.1: *Tabella mendeliana: la prima idea di genetica.*
Fonte "Introduction to heredity: Mendel and his peas" di Khan [21]

1952: informazione genetica nel DNA

Prima di questo anno si pensava che l'informazione genetica venisse portata dai complessi proteici. A smentire ciò e a ad avere l'intuizione che in realtà fosse il DNA sono stati Alfred Hersey e Martha Chase, con un esperimento che prende appunto il loro nome. L'esperimento Harsey-Chase prevedeva di infettare dei batteri di Escherichia coli con virus. Il procedimento prevedeva l'utilizzo di due culture di E.coli, una marcata radioattivamente nei nucleotidi e l'altra nei complessi proteici di cisteina. In seguito le due culture venivano infettate da virus, i fagi T2, che si replicavano e davano origine a virus marcati internamente, nel caso di contatto con la cultura DNA-marcata, e gli altri esternamente, nel caso di interazione con i proteine-marcati. Ma i virus per poter infettare la cellula ospite perdono la parte esterna per poter entrare al suo interno, pertanto i due biologi dedussero così che l'informazione genetica era contenuta nel DNA e non nelle proteine [16].

1953: il dogma centrale

Rosalind Franklin riesce per la prima volta nella storia a osservare la struttura a doppia elica del DNA con l'utilizzo dei raggi-X, facendo diventare la sua *Photo 51*, vedi Figura 2.2, una delle fotografie scientifiche più conosciute e stampate nella storia. [49]
Come tutti però sappiamo, il suo riconoscimento con premio Nobel avverrà solo dopo la sua morte, e il riconoscimento immediato fu dato a Watson e Crick, che riescono ad sviluppare il modello matematico per la diffrazione del DNA, studio già iniziato dalla Franklin stessa, e a rappresentare la doppia elica con un modello osservabile da tutti.
La disposizione delle basi azotate all'interno dell'acido desossiribonucleico diventa di dominio pubblico, le coppie Adenina-Timina e Citosina-Guanina diventano dunque le responsabili della ritenzione e del passaggio delle informazioni genetiche.

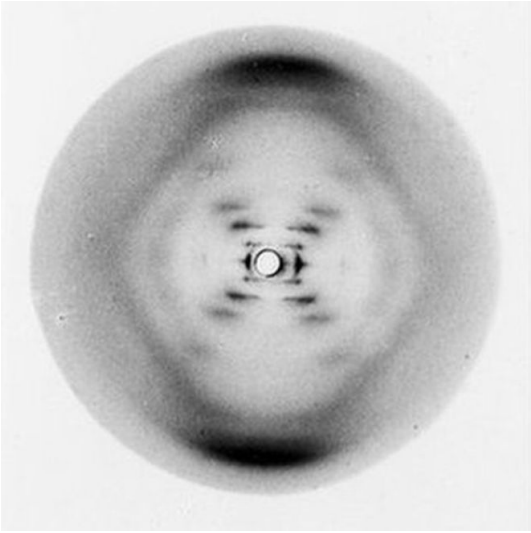


Figura 2.2: *Prima visione della doppia elica del DNA attraverso raggi-X eseguita da Rosalind Franklin.*
Fonte "The most important photo ever taken?" di Fergus Walsh [48]

1955: primo sequenziamento

Seppur ci sia stata la grande scoperta del DNA che ha portato a stravolgere tutte le idee sulla biologia, il biochimico Frederik Sanger ha proceduto ugualmente a sequenziare per la prima volta i complessi proteici dell'insulina bovina [31]. Il motivo per cui le proteine sono state preferite al DNA è banalmente la tecnologia e disposizione a quel tempo. Per questa pubblicazione è stato utilizzato la degradazione di Edman, che prevedeva il sequenziamento degli amminoacidi all'interno dei peptidi. Questo metodo però permetteva di sequenziare frammenti di solo 50-60 amminoacidi, pertanto per proteine più grandi si doveva fare uno sforzo inimmaginabile per riunire i frammenti e tutto ciò solo attraverso l'osservazione dell'occhio umano.

1960: la pioniera della bioinformatica

Margaret Dayhoff è definita la madre e il padre della bioinformatica poiché ha unito per la prima volta la biologia, la medicina e l'informatica, tre competenze che possedeva lei stessa.
Con la collaborazione di Robert Landley, la Dayhoff ha progettato *COMPROTEIN*, vedi Figura 2.3, un programma designato per scovare la struttura primaria delle proteine usando la *degradazione di Edman* e sfruttando il nuovo mainframe IBM 7090 dell'università su cui aveva scritto il programma con il linguaggio di programmazione *FORTRAN*, con il compito di ricevere in input diverse sequenze di frammenti e restituire una sola sequenza rappresentante la catena di amminoacidi della proteina [26].

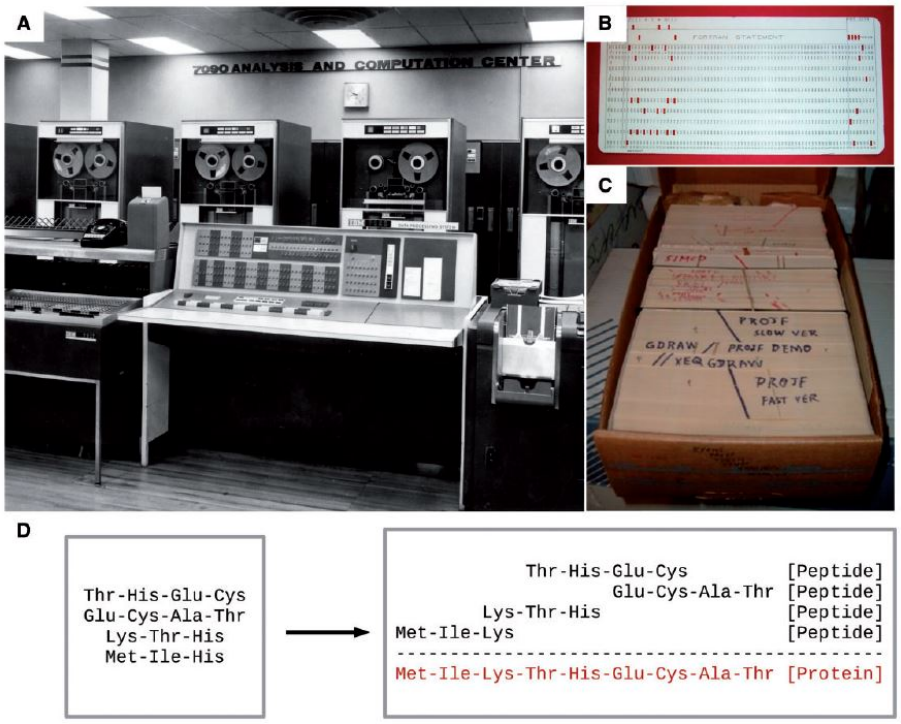


Figura 2.3: *Il primo software bioinformatico COMPROTEIN composto da: (A) l'IBM 7090, (B) codice fortran, (C) intero listato del codice, (D) esempio di input e output.*
Fonte "A brief history of bioinformatics" di Gauthier et al. [26]

Questo software prevedeva dunque di poter allineare i vari frammenti per ottenere così il sequenziamento totalitario anche di una grande proteina, risolvendo così il problema degli anni precedenti. Inizialmente nel COMPROTEIN ogni amminoacido aveva un identificativo di tre lettere, in seguito però la Dayhoff ha sviluppato una cifratura a una sola lettera per semplificarne lo studio e diminuite i tempi computazionali. Questa cifratura a un simbolo alfabetico è usata tutt'oggi [20].
Dopo aver fatto ciò, ha anche pensato di creare un enorme database accessibile per permettere a tutti i ricercatori di fare interconnessioni tra le varie proteine di differenti organismi, e dunque arrivare a delineare la storia dell'evoluzione delle specie. Nel 1965 la Dayhoff pubblica sulla rivista *Atlas of Protein Sequence and Structure* il sequenziamento di 65 proteine da lei fatto, andando così a dare origine al primo database bioinformatico, seppur cartaceo, e dando un'enorme spinta allo sviluppo di questo nuovo campo di ricerca [20].

1970: la bioinformatica al servizio della paleogenetica

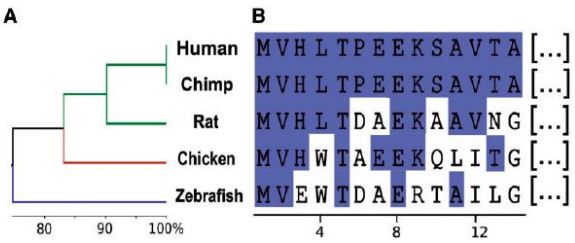


Figura 2.4: *Similitudini tra proteine ortologhe (B) in diversi organismi (A).*
Fonte "A brief history of bioinformatics" di Gauthier et al. [26]

L'utilità di confrontare i complessi proteici di diversi organismi viventi per risalire ai loro predecessori e scoprire che magari appartenevano allo stesso filone evolutivo ha fatto un enorme passo avanti con la ideazione di un software automatico, [20]. Needleman e Wunsch hanno pensato in questo anno a sviluppare un programma dinamico che permettesse di confrontare diverse sequenze proteiche e nucleotidiche di diversi organismi, vedi Figura 2.4, venendo così in aiuto dei paleogenetisti che prima di ciò dovevano arrivare alle conclusioni semplicemente osservando la similarità. Solo un decennio dopo però verrà pubblicato il loro software di global alignment.

1978: il PAM

La Dayhoff, stavolta affiancata da Schwartz e Orcutt, ha messo un'altra sua firma nelle grandi invenzioni della bioinformatica andando a sviluppare un modello probabilistico sulle mutazioni degli amminoacidi all'interno della catena proteica. Nasce così il Point Accepted Mutation, PAM, che stabilisce la possibilità di una mutazione di un amminoacido all'interno della struttura primaria della proteina senza che quest'ultima vada a originare un impatto nel processo di selezione naturale. Il PAM viene inoltre usato come misura dell'evoluzione delle specie [20].

1979: ritorno al DNA

Seppur siano passato diversi decenni dalla constatazione che il DNA fosse il vero portatore dell'informazione genica, solo a fine degli anni 70 si è riusciti ad avere accesso alle informazioni relative ai geni di un intero organismo. Nel 1977 è stato sequenziato il primo genoma che apparteneva a un batteriofago e che ha sfruttato il criterio dei terminatori di catena o metodo di Sanger [20]. Nel 1979 questa tecnica è stata ampliata con un software dedicato alla sua analisi.

1984: primo pacchetto

In questo anno nasce il primo programma che poteva essere installato all'interno di un minicomputer. Il *WWW2GCG*, così chiamato, era un pacchetto contenente 33 funzioni per manipolare il DNA, l'RNA o le proteine, con versioni per linguaggi *FORTRAN*, *Perl* e *C* rendendolo così molto versatile [6].

1985: la rivista ufficiale

Nasce in questo anno la rivista *Computer Applications in the Biosciences*, CABIOS, che diventerà tre anni dopo *Bioinformatics* e avrà anche una copia online. La rivista sorge con lo scopo di sensibilizzare i biologi all'utilizzo dell'informatica per i loro studi anche con l'organizzazione di conferenze [30].

1987: standards

Durante questi anni diverse organizzazioni di database scientifici decidono di unirsi per facilitare lo scambio di dati ma anche per standardizzare le diverse nomenclature. Nascono così l'*International Nucleotide Sequence Database Collaboration*, unione dell'EMBL (European Molecular Biology Laboratory), GenBank e della DDBJ (DNA Data Bank of Japan) [20].

1990: World Wide Web

Oramai non riusciamo a immaginarci un mondo senza il world wide web, senza le informazioni a portata di click, senza uno scambio immediato di messaggi e dati; ma se dovessimo ragionarci anche solo per un attimo realizzeremmo che c'è stato un periodo non tanto lontano in cui tutti gli scambi e le consultazioni di informazioni erano fisiche, con dischi magnetici o con fogli cartacei, rendendo lungo, difficile e tedioso uno studio innovativo. Con la nascita del *World Wide Web* sviluppato dal *CERN*, tutte le informazioni sono divenute di dominio pubblico e di immediato accesso, e così è avvenuto anche per i dati bioinformatici. Nasce così una collaborazione mondiale tra ricercatori, con diffusione di scoperte e correzione di eventuali errori di trascrizione, rendendo sempre più fedeli le informazioni. Nascono anche nuovi database come *Genomes*, *PubMed* e *Human Genome* [24].
Nello stesso anno Python è diventato il linguaggio più diffuso tra i bioinformatici, soprattutto grazie al suo tool di allineamento locale *BLAST (Basic Local Alignment Search Tool)* [20].

2000: biologia dei sistemi

Con il progetto *Human Genome* sono riusciti ad identificare i geni appartenenti alla catena di DNA, ma ancora si sapeva poco di come questi interagissero tra loro [39]. Per questo motivo agli inizi del secondo millennio si decide di spostare l'attenzione verso un approccio più dinamico, indagando le relazioni tra i vari componenti e le loro mutazioni: la *biologia dei sistemi*, vedi Figura 2.5.

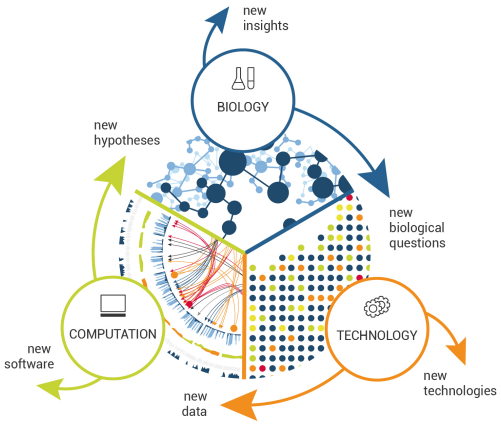


Figura 2.5: *Schema sul funzionamento generale della biologia dei sistemi. Fonte "Disease and systems biology" di omicsout [37]*

Biologia dei sistemi

La moltitudine di dati affiorati in questi pochissimi anni ha portato alla nascita di numerosi siti che permettessero la loro visione più dinamica sotto forma di reti e non unicamente statica come avveniva nei decenni precedenti.
Non si va a ricercare solo la singola entità, ma anche la sua interconnessione con altri elementi. Questo modo di rappresentare i dati è molto utile per i biologi ricercatori, i quali facendo esperimenti, su topi knocked-out ad esempio, osservano la presenza di attività su più geni o elementi biologici. La ricerca delle motivazioni di queste misure si fa via via sempre più difficoltosa, dovuto soprattutto all'evoluzione della tecnologia a disposizione, pertanto uno strumento del genere risulta molto utile e di beneficio.

Con il *The Human Genome Project* si è arrivati ad avere le sequenze di tutti i geni presenti nella catena di DNA umano, vedi Figura 2.6, e dunque la biologia si è spostata su uno studio più dinamico di tutti questi dati, vedi Figura 2.7. Gli studi con una visione più dinamica prevedono l'investigazione delle relazioni che accomunano due o più entità, che può andare da una fosforilazione a una repressione dovuta all'azione di un micro RNA che si aggancia al sito di un determinato gene.

Tutti questi dati vengono raccolti sperimentalmente durante studi in laboratorio, riportati su articoli scientifici e talvolta organizzati in database online che evidenziano le relazioni appena scovate. Questi siti permettono dunque di raccogliere tutte le relazioni tra le varie entità e di evidenziarle in modo chiaro e facilmente interpretabile al biologo che affronta degli studi simili.
In alcuni casi vengono mostrate anche relazioni statistiche, cioè connessioni non riscontrate sperimentalmente ma che un algoritmo informatico intravede e prevede. In altre parole, se un elemento interagisce notevolmente con altri elementi e questi ultimi hanno una relazione con un altro elemento, statisticamente si prevede un nesso tra i due. Un esempio delle relazioni statistiche sono quelle di co-espressione.

Esistono dunque due tipi di *system biology databases*: uno basato su evidenze scientifiche, il quale rappresenta la maggior parte dei siti in circolazione in quanto necessita di prelevare unicamente le informazioni dagli articoli pubblicati; e l'altro su considerazioni statistiche [5]. Unendo i due si ha una visione più ampia della biologia molecolare.

Gli elementi, precedentemente sequenziati per intero e di cui sono state scoperte le relazioni, vengono riportati sotto forma di grafi, cioè nodi collegati tra loro da archi. I nodi che rappresentano le entità sono uniti attraverso segmenti direzionati che identificano la relazione biologica che li accomuna.
I grafi, comunemente conosciuti come reti, sono uno strumento molto utile per gli studi informatici, in quanto si possono applicare algoritmi di risoluzione anche propri di altre discipline come le telecomunicazioni.
Attraverso questo potente strumento si possono condurre diversi studi sugli elementi che nella pratica sperimentale si riscontrano, ad esempio si ha conferma di un risultato, oppure si può cercare di capire che problemi ha causato una mutazione.

Figura 2.6: *Esempio di analisi statica di un elemento biologico: sequenza di basi azotate. Fonte "The code of life" di Hunter [19]*

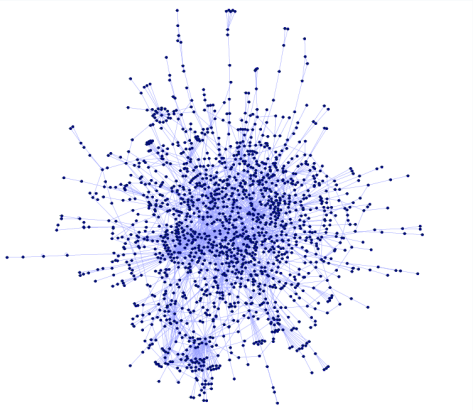


Figura 2.7: *Esempio di analisi dinamica di elementi biologici: grafo con PPI di un lievito. Fonte "Machine learning for pairwise data : applications for preference learning and supervised network inference" di Birlutiu [4]*

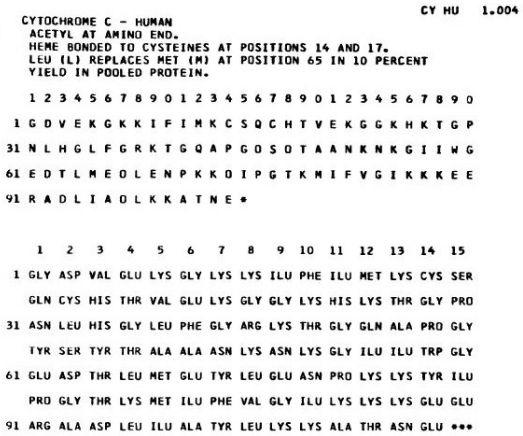
Capitolo 3

The RING database

Nel 2018 la rivista *Nucleic Acids Research* ha identificato la presenza di 1737 diversi database di dati bioinformatici [7].
Si possono suddividere in diverse macro-categorie[32]:

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
- Cell biology

Come si può osservare, dal 1965 a oggi si è passati dal pubblicare il primo database bioinformatico in forma cartacea, esempio in Figura 3.1, al poter ricercare rapidamente in formato elettronico milioni di interazioni tra diverse entità biologiche contenute in una moltitudine di database, esempio in Figura 3.2.



CY HU 1.004

CYTOCHROME C - HUMAN
ACETYL AT AMINO END.
HEME BONDED TO CYSTEINES AT POSITIONS 14 AND 17.
LEU (L) REPLACES MET (M) AT POSITION 65 IN 10 PERCENT
YIELD IN POOLED PROTEIN.

```
   1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
 1 G D V E K G K K I F I M K C S Q C H T V E K G G K H K T G P
31 N L H G L F G R K T G Q A P G F S Y T A A N K N K G I I W G
61 E D T L M E Y L E N P K K Y I P G T K M I F V G I K K K E E
91 R A D L I A Y L K K A T N E *

    1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
 1 GLY ASP VAL GLU LYS GLY LYS LYS ILU PHE ILU MET LYS CYS SER
   GLN CYS HIS THR VAL GLU LYS GLY GLY LYS HIS LYS THR GLY PRO
31 ASN LEU HIS GLY LEU PHE GLY ARG LYS THR GLY GLN ALA PRO GLY
   TYR SER TYR THR ALA ALA ASN LYS ASN LYS GLY ILU ILU TRP GLY
61 GLU ASP THR LEU MET GLU TYR LEU GLU ASN PRO LYS LYS TYR ILU
   PRO GLY THR LYS MET ILU PHE VAL GLY ILU LYS LYS LYS GLU GLU
91 ARG ALA ASP LEU ILU ALA TYR LEU LYS LYS ALA THR ASN GLU ***
```

Figura 3.1: *Prima sequenza del citocromo C umano. Fonte "ATLAS of PROTEIN SEQUENCE and STRUCTURE 1965" di Dayhoff et al. [8]*

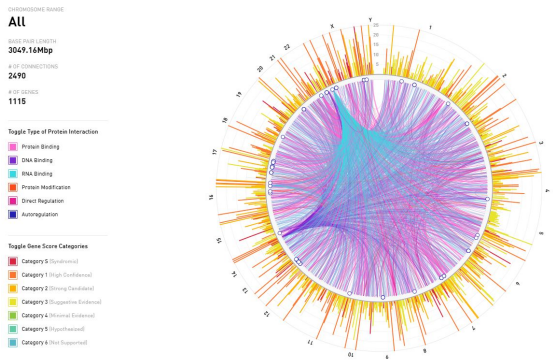


Figura 3.2: *Geni e PPI di tutti i cromosomi umani rappresentati sul database SFARI. Fonte "SFARI database" di The Simons Foundation [10]*

Tutto ciò è particolarmente positivo se si pensa all'evoluzione della bioinformatica e di come questa ha destato interesse in molti, ma d'altra parte c'è la confusione che si crea con tutta questa scelta. Molto preoccupante è anche la non esistenza di standard internazionali che permettono una lettura e un utilizzo facilitato dei dati, portando dunque ad ancora più problematiche. Ma non solo, gli stessi database avevano la tendenza a non aggiornare i propri dati e addirittura a conferire le stesse nomenclature a dei nuovi.

Per ovviare a questi problemi è nato il *The RING Database*, una piattaforma sviluppata dai professori ricercatori del Politecnico di Torino: Alfredo Benso, Gianmarco Politano e Stefano Di Carlo, con l'intento di uniformare e convogliare a sè diverse informazioni provenienti da più database, per citarne alcuni:
DGIdb, con interazioni gene-farmaco;
DisGeNET, racchiudente associazioni gene-malattia e varianti di quest'ultima;
DrugBank, avente le associazioni farmaco-target;
FisinGene, il quale racchiude 8 database aventi interazioni proteina-proteina;
KEGG, l'enciclopedia di geni e genomi di Kyoto;
tf2dna, unione di 10 database di transcription factors;
TargetScan, con interazioni miRNA-target. Come detto in precedenza lo scopo di questo nuovo database non è solo lo sviluppo di un enorme complesso di dati, aspetto da non sottovalutare in quanto è molto utile per rendere i dati sempre più affidabili, ma è anche un mezzo di standardizzazione.

Il metodo di uniformazione dei dati attraverso standard è visibile con un semplice tocco sull'entità; ad esempio, nel caso degli oggetti *gene* e *transcription factor* i campi visibili sono: *type, ID, name, isTF, uniprot_id, alias, map_location, uniprot, genexpatlas, ncbi*, Vedi Figure 3.3a e 3.3b. In particolare l'univocità della nomenclatura è appunto osservabile nel campo *ID*, mentre l'*alias* risulta essere un campo importante per ritrovare le corrispondenze in altri database.
Per quanto riguarda i micro RNA i campi rappresentati sono *type, ID, name, extra info*, vedi Figura 3.3c; l'ultimo campo solitamente riporta l'ID del precursore.

TYPE	ID	NAME	isTF	uniprot_id	alias	map_location	uniprot	genexpatlas	ncbi
gene	11011	TLK2	0	B4DUC2	HsHPK PKU-ALPHA	17q23.2	UNIPROT	GeneExpAtls	NCBi

(a) *Gene TLK2*

TYPE	ID	NAME	isTF	uniprot_id	alias	map_location	uniprot	genexpatlas	ncbi
gene	55723	ASF1B	1	A0A024R7G4	CIA-II	19p13.12	UNIPROT	GeneExpAtls	NCBi

(b) *Transcription factor ASF1B*

TYPE	ID	NAME	EXTRA INFO
miRNA	MIMAT0014998	hsa-miR-3133	Precursor Id: MI0014153

(c) *micro-RNA hsa-miR-3133*

Figura 3.3: *Esempi di standardizzazione applicate dal The RING database [12]*

Con l'utilizzo di una nomenclatura ben definita che accomuna tutte le entità e una grafica gradevole alla vista e semplice, lo studio e la ricerca di interazioni diventa immediatamente meno tediosa e più agevole, riducendo in questo modo anche gli errori di lettura e confronto di dati con terzi. La sua architettura è progettata su diversi layer caratterizzati da un aumento di informazioni passando da uno all'altro. Partendo dal più basso si ha [12]:

- Raw Data Layer
  Contiene le interazioni tra le varie entità biologiche e le loro specifiche, come le informazioni sui geni con le loro posizioni all'interno del cromosoma; i transcription factors (TF) e i co-transcription factors; le interazioni proteina-proteina (PPI); l'ontogenesi dei miRNA e i loro target; i polimorfismi a singoli nucleotidi (SNP); i farmaci e infine le malattie.

- Omics Layer
  Si può definire lo strato dei standard, infatti riorganizza i dati e gli da un ID specifico per permetterne una facile lettura. Rinominando le entità non si perdono però le etichette originarie e il database di provenienza, informazioni accessibili in ogni momento dello studio.

- Model Layer
  L'interfaccia utente con la possibilità di organizzazione della rete risultante in tabelle o grafi facilmente manipolabili e leggibili, anche grazie al diverso codice colore identificativo di ogni entità.

La struttura così composta permette pertanto di ottenere ciò che uno standard internazionale avrebbe già dovuto fare, ma che ancora non esiste forse per via della diffusione recente di questo ambito. Il database è fruibile sul web a tutti, anche da chi non ha così padronanza degli strumenti informatici [12].
All'interno del sito basta una semplice ricerca con i nomi delle entità per accedere alle loro informazioni sui database da cui sono state prese e la loro nomenclatura. Inoltre è possibile osservare queste entità interconnesse tra loro attraverso degli archi rappresentativi delle varie funzioni chimico-biologiche.
Per ottenere un grafo facilmente consultabile e di gradevole aspetto, è prima necessario settare dei filtri che permettono di scegliere i dataset desiderati; in particolare è possibile selezionare/deselezionare diversi tipi di entità (geni, transcription factors, ecc.) e anche se si vogliono o meno le entità e gli archi validati o curati manualmente oppure direzionali.

Dopo aver fatto questo primo passaggio, si passa alla ricerca delle entità attraverso l'inserimento dei nomi nell'apposita barra, vedi Figura 3.4.

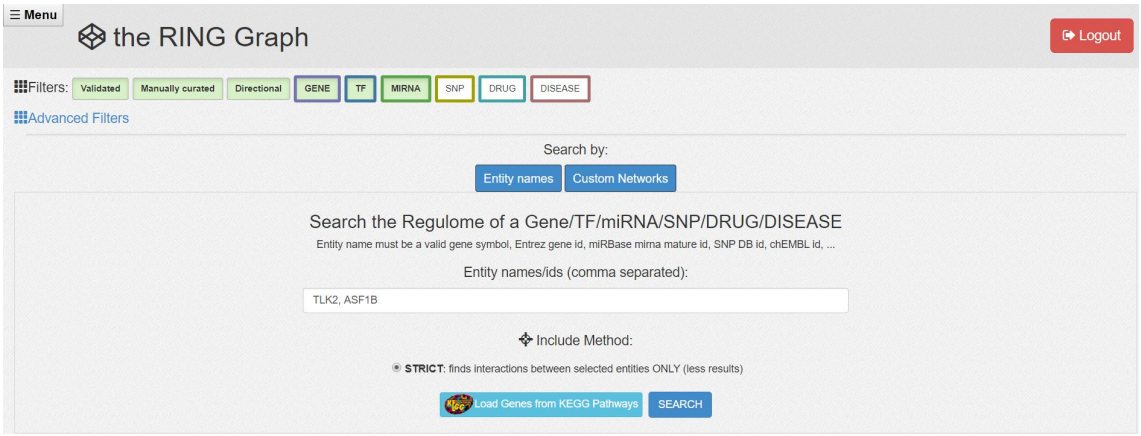


Figura 3.4: *Schermata iniziale del The RING database [12] dove è possibile ricercare le entità e imporre alcuni filtri*

Non appena il tasto *search* viene calcato, tutte le entità compaiono nella finestra di visualizzazione sottostante.
In seguito si può pensare ad espandere le entità introducendo altri nodi e interconnessioni, scegliendo anche il grado di complessità ed espansione, come il network in Figura 3.5 che rappresenta le interazioni, comprendenti i miRNA, tra il gene TLK2 e il transcription factor ASF1B.

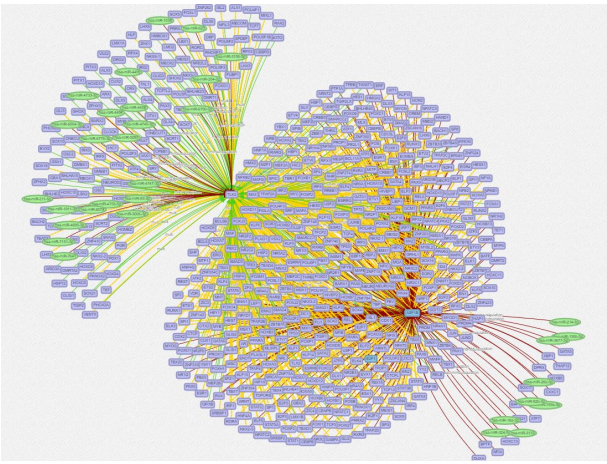


Figura 3.5: *Esempio di grafo ottenuto con il The RING database [12] contenente elementi appartenenti al gene TLK2 e al transcription factor ASF1B*

Tutto ciò è una consultazione online che si perde non appena si esce dal sito, pertanto se l'utente ha intenzione di continuare il suo studio con quel grafico è bene che esporti le sue proprietà. I file che riassumono gli oggetti, chiamati nodi, e le loro relazioni, chiamate archi o edges, sono: l'immagine, un documento excel, un file sif, oppure i due documenti di nodi e file in formato .csv.

I file dei nodi e degli archi

I file scaricabili dal sito *nodes.csv* e *edges.csv* sono risultati molto utili nello studio dei grafi attraverso l'utilizzo di un ambiente di sviluppo che sfrutta il linguaggio Python. Analizzando i due file, facilmente consultabili con Excel, è possibile capire come il The RING Database ha introdotto una standardizzazione facile ed efficace sia per i nodi che per le relazioni tra essi.

Nel file **nodes.csv** si trova una descrizione di ogni nodo, quindi di ogni entità biologica, come si può osservare in Figura 3.6; in particolare le caratteristiche descritte in questo file sono:

- *Type*
  Il "tipo" di entità, cioè se si tratta di un gene, una proteina, un transcription factor, un miRNA, un polimorfismo a singolo nucleotide, una malattia oppure un farmaco.
- *ID*
  L'ID, cioè l'identificativo per cui è stato memorizzato nel database da cui è stato prelevato.
- *Name*
  Il nome dell'entità dipende dal database di origine, come per l'id, per permettere una facile ricerca in quest'ultimo.
- *Extra info*
  Le info aggiuntive possono essere: altri id appartenenti ad altri database, come l'UniProt, il più grande database di proteine appartenenti agli organismi viventi e virus; aliases, cioè altri nomi con cui queste entità sono state etichettate in altri database; e infine la collocazione all'interno della catena.
  Per quanto riguarda i miRNA le informazioni in più sono l'identificativo del precursore.

TYPE	ID	NAME	EXTRA INFO
gene	11011	TLK2	UniprotId: B4DUC2, Aliases: HsHPK|PKU-ALPHA, Map L...
TF	55723	ASF1B	UniprotId: A0A024R7G4, Aliases: CIA-II, Map Location:...
TF	1869	E2F1	UniprotId: Q01094, Aliases: E2F-1|RBAP1|RBBP3|RBP3...

Figura 3.6: *Estratto del file nodes.csv di 4 righe*

Nel file **edges.csv** invece si trovano sia le informazioni sui due nodi interagenti, sia le caratteristiche della relazione che li lega, vedi Figura 3.7; più precisamente:

- *Source ID/name/type*
  L'ID/nome/tipo del nodo sorgente, cioè da cui diparte l'arco.

- *Target ID/name/type*
  L'ID/nome/tipo del nodo target, cioè a cui arriva l'arco.

- *Direction*
  In questa colonna sono contenuti diversi simboli che rappresentano l'azione dell'arco. Può essere di inibizione e dunque terminare con '—|'; di attivazione con il simbolo '—>'; di co-attivazione '<.—.>', spesso usato per due entità che si attivano a vicenda dando origine a un complesso d'attivazione; e infine di relazione a livello di proteoma, caratterizzata dal simbolo '—'.

- *Action*
  Descrive il significato dell'arco in termini più biologici, come ad esempio la fosforilazione, il binding, l'affinità con la tecnologia cromatografica (dunque un possibile errore di lettura), repressione ed altri.

- *db*
  Questa colonna riporta i database da cui è stata prelevata l'informazione riguardante la relazione tra le due entità.

- *PMID*
  Infine si trova l'informazione riguardante l'identificativo delle pubblicazioni su PubMed che riportano questa relazione.

source_id	source_name	source_type	target_id	target_name	target_type	direction	action	db	PMID
1869	E2F1	tf	*55723*	ASF1B	gene	t->	Activation	TRRUST	*17328667*
1869	E2F1	gene	*55723*	ASF1B	gene	t-|		Pscan-Kulakovskiy2013	*-1*
1869	E2F1	gene	*55723*	ASF1B	gene	t-|		Pscan-Matys2006	*-1*
1869	E2F1	gene	*55723*	ASF1B	gene	t-|		Pscan-Mathelier2014	*-1*

Figura 3.7: *Estratto del file edges.csv di 4 righe*

Capitolo 4

Teoria dei grafi

Sono molte le situazioni che possono essere studiate mediante l'utilizzo di grafi; infatti, frequentemente ci si trova ad affrontare problemi nei quali devono essere definite le relazioni che intercorrono tra diversi oggetti, sia in termini qualitativi che quantitavi, rendendo possibile la modellazione del sistema mediante un grafo, il quale può assumere diverse peculiarità a seconda del contesto.
Un grafo, come riportato in [11], è rappresentato da un insieme non vuoto e finito di elementi, chiamati vertici o nodi, congiuntamente ad un set di coppie di vertici non ordinate, chiamati archi, i quali uniscono i due vertici, rendendoli adiacenti. Ogni arco collegato ad un determinato nodo risulta essere incidente con il nodo stesso, mentre gli archi che condividono un nodo risultano essere adiacenti. Il grafo può quindi essere definito graficamente mediante punti e segmenti, che rappresentano rispettivamente archi e nodi, come mostrato in Figura 4.1. Esso, però, può anche essere rappresentato mediante matrici, come nel caso della matrice di adiacenza, la quale è una matrice quadrata, usata spesso per estrarre le informazioni dalla rete mediante tecniche computazionali, nella quale il numero di righe e colonne è pari al numero di nodi, mentre gli elementi che la compongono possono assumere solo due stati:

- 1, se l'arco congiungente i nodi rappresentanti riga e colonna della matrice esiste
- 0, altrimenti

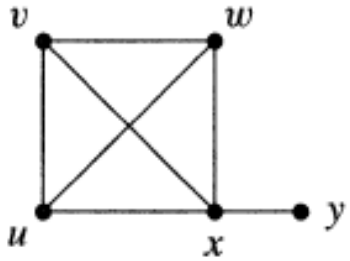


Figura 4.1: *Grafo non direzionato.*
Fonte "Introduction to Graph Theory-second edition" di West [3]

Un grafo può essere anche direzionato ed essere rappresentato da una rete nella quale le coppie di vertici, rappresentanti gli archi, sono ordinate, in modo tale da ottenere

come primo elemento la sorgente e come secondo il target. Nella rappresentazione grafica di un grafo direzionato, gli archi non sono più semplici segmenti, bensì frecce che ne rappresentano la direzione, come mostrato in Figura 4.2.

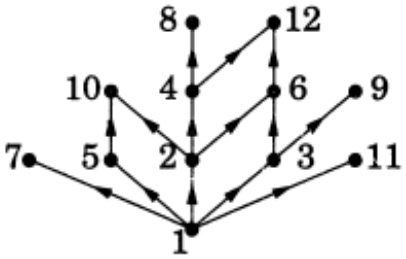
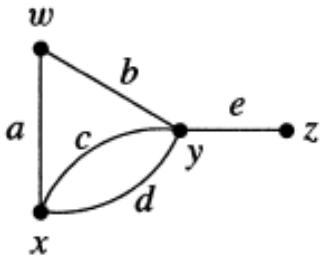
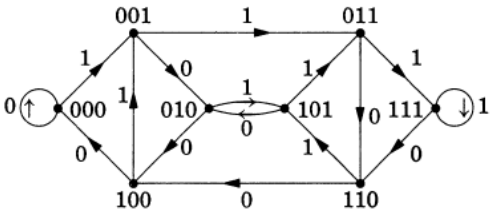


Figura 4.2: *Grafo direzionato.*
Fonte "Introduction to Graph Theory-second edition" di West [3]

Volendo dare un'ulteriore differenziazione dei grafi, possono essere detti multigrafi quei grafi che consentono di avere archi multipli tra 2 sorgenti; analogamente, esistono i multigrafi direzionati. Ovviamente, i diversi archi che coesistono tra le stesse sorgenti, portano diverse informazioni, che quindi ne rappresentano la principale distinzione. Esempi di rappresentazione di multigrafo non direzionato e direzionato sono mostrati in Figura 4.3.



(a) *Multigrafo non direzionato*



(b) *Multigrafo direzionato*

Figura 4.3: *Grafi con archi multipli.*
Fonte "Introduction to Graph Theory-second edition" di West [3].

I nodi possono avere diverse proprietà, come il grado di connettività, rappresentato dal numero di archi incidenti in ogni nodo, diversi tipi di centralità, o la posizione e il coinvolgimento nei diversi processi. Analogamente, anche gli archi possono avere diverse caratteristiche, come diversi tipi di centralità, gli attributi riguardanti la relazione caratterizzante l'arco, o la direzionalità, se presente. Ognuno di questi attributi può contribuire ulteriormente ad analizzare la topologia della rete [2], la quale permette non solo di stabilire l'importanza e il valore della rete stessa nel contesto analizzato, ma anche di estrarre informazioni più approfondite.

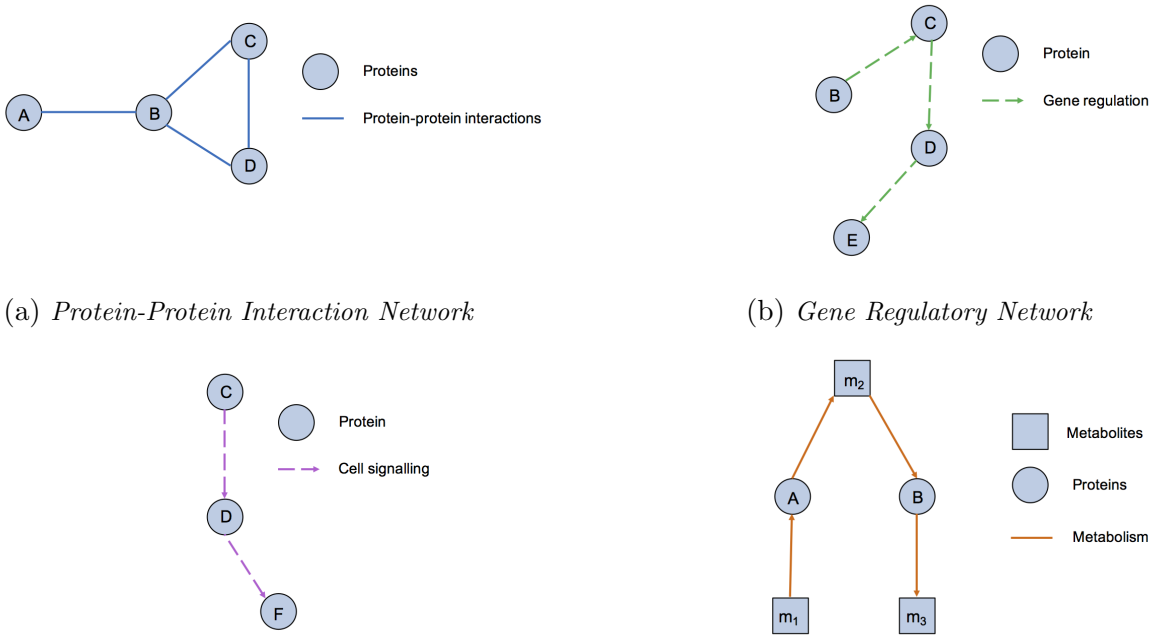
Una rete può essere:

- **Scale-free**: rete nella quale la distribuzione di probabilità che un vertice, scelto in maniera random, abbia un certo numero di legami, è di tipo power-law, anche variando il numero di nodi presenti in essa.
- **Assortative**: nella quale i nodi con un alto numero di legami (un alto degree), tendono a formare legami con altri nodi caratterizzati dalla stessa proprietà.
- **Dissortative**: nella quale i nodi con un alto numero di legami, tendono a collegarsi a nodi con basso numero di legami [23].
- **Small-world**: rete nella si possono identificare un gran numero di gruppi, o cluster, uniti da piccoli collegamenti [2].

La teoria dei grafi rappresenta, come già detto, uno strumento utile in diverse discipline, compresa la biologia. Infatti, per analizzare un problema di tipo biologico, molto spesso non ci si può ridurre all'analisi dei singoli componenti, ma bisogna analizzare le diverse interazioni tra loro, come nel caso del drug targeting, della regolazione dei diversi geni, o della diagnosi precoce di determinate patologie. Nel campo della Systems Biology, la teoria dei grafi è stata applicata con successo per via della sua flessibilità e scalabilità, sia al fine di ottenere una visione globale, scale-free, del sistema, sia per ottenere una visione ravvicinata ed identificare motivi ricorrenti e interazioni specifiche tra diverse entità. Le principali reti di interesse, in campo biologico e medico sono essenzialmente 4, come riportato in [23]:

- **Protein-Protein Interaction (PPI) Network**, mostrato in Figura 4.4a, solitamente rappresentati da grafi semplici, che rappresentano la coordinazione e la cooperazione delle diverse proteine, finalizzati allo studio della funzione delle diverse proteine, rappresentante una grande incognita nel mondo della biologia. Attraverso diverse tecniche, è stato possibile ricavare una stima dell'interazione delle diverse proteine all'interno degli organismi, rendendo infatti disponibili diversi database, contenenti informazioni riguardo i dati delle PPI. Queste reti risultano essere di difficile interpretazione, a causa delle diverse modalità di estrazione dei dati da processare. In queste reti, i componenti assumono un'importanza biologica direttamente proporzionale al loro tasso di connettività.
- **Gene Regulatory Netwkorks (GRNs)**, mostrato in Figura 4.4b, solitamente rappresentati da grafi direzionati, che contengono informazioni riguardanti l'espressione genica e il suo controllo, coadiuvato da diversi fattori. La direzionalità permette di ottenere un'indicazione e una modellazione degli eventi coinvolti nell'espressione genica. Anche esse sono caratterizzate, nella maggiorparte dei casi, da una topologia di tipo scale-free. Sono inoltre molto sensibili all'evoluzione, mutandone continuamente le dinamiche.

- **Signal Transduction Network**, mostrato in Figura 4.4c rappresentati da multigrafi in grado di simulare le diverse interazioni tra le bioentità, e investigarne la trasmissione dei segnali. In base alle diverse circostanze e ai parametri ambientali, si possono generare risposte diverse. Anche esse sono Scale-free.
- **Metabolic and Biochemical Network**, mostrato in Figura 4.4d, utilizzati per studiare percorsi metabolici, contenenti informazioni su eventi di tipo biochimico e la loro correlazione, la cui unione permette di costituire una rete metabolica completa. Sono sia scale-free che small-world e subiscono numerose variazioni, variando l'organismo analizzato, pur mantenendo il diametro della rete quasi costante.



(a) *Protein-Protein Interaction Network*

(b) *Gene Regulatory Network*

(c) *Signal Transduction Network*

(d) *Metabolic and Biochemical Network*

Figura 4.4: *Quattro tipi di network biologici.*
Fonte "Types of biological networks" di Train online [36]

Bisogna segnalare che in queste reti, possono essere presenti dei pattern ricorrenti, chiamati Network motifs, composti da pochi nodi e dagli archi di collegamento tra essi, i quali risultano avere una grande importanza, a causa della loro influenza sulle dinamiche del sistema totale [2]. La rete sulla quale è stato condotto il nostro lavoro di tesi è estratta dal *The RING Database* [12], il quale, come detto nel Capitolo 3, si prefigge lo scopo di standardizzare ed unificare diverse informazioni, provenienti da diverse risorse e appartenenti a diverse sfere, ottenendo quindi una rete eterogenea che riesce ad unificare diversi livelli biologici, non configurandosi esclusivamente in nessuna delle reti illustrate precedentemente, bensì rappresentandone un connubio. Infatti, sia gli interattori che le relazioni intercorrenti tra essi, che rappresentano il focus del progetto, sono di diverso tipo e appartengono a categorie diverse.

Ogni entità ed ogni interazione è stata standardizzata, di modo tale da restituire un identificativo univoco. Le entità hanno diversi attributi, che ne facilitano l'inquadramento biologico, come la direzione, appartenente ad un dizionario di simboli in grado di rappresentare il significato dell'interazione biologica analizzata; l'azione, riportante l'annotazione originale utilizzata per definire il simbolo direzionale corrispondente, di modo tale da poter utilizzare lo stesso simbolo per meccanismi d'azione diversi e differenziarle tra loro; lo score, che restituisce il livello di confidenza ereditato dal database di origine. Tutto ciò fa si che la rete sul quale sono stati condotti gli studi, non abbia un inquadramento preciso, ma sia un'unica rappresentazione multilivello di ogni interazione possibile. Tutte le reti sono state elaborate usando il linguaggio di programmazione ad alto livello Python, orientato agli oggetti, in entrambe le versioni 2.7 e 3.8. Le reti sono state estratte dal TheRingDB sia in formato .csv, che come oggetto serializzato di tipo *.gpickle*.
Sono state sfruttate essenzialmente 2 librerie per l'elaborazione dei dati:

- **NetworkX**: la libreria fornisce classi e metodi specifici per creare, estrarre ed elaborare oggetti di tipo "grafo" [14].

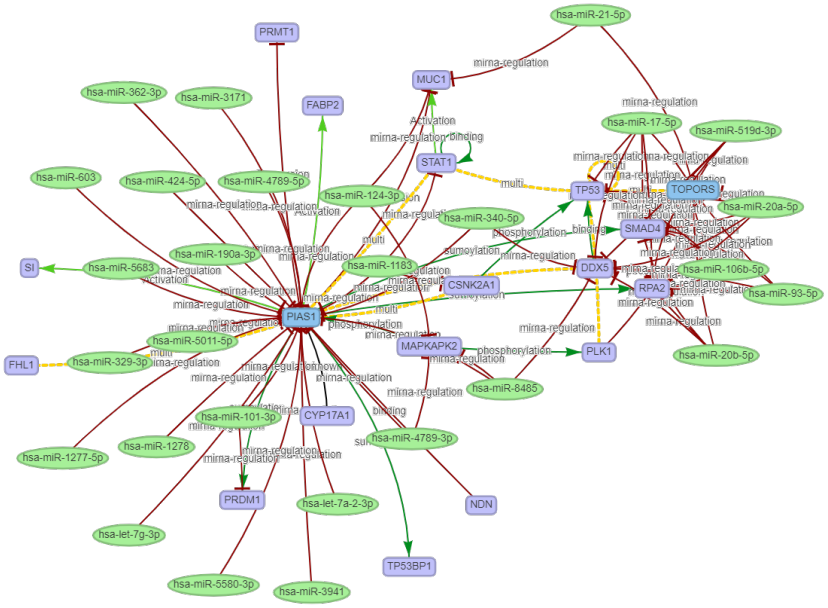
- **Pandas**: la libreria fornisce metodi per implementare strutture di dati flessibili, in modo tale da ottenere uno strumento pratico e potente per l'analisi di dati effettuata in ambiente Python [25].
  Per la rappresentazione grafica delle reti, dove richiesta o implementata, è stata utilizzata la libreria Matplotlib, che ha permesso di rappresentare la rete usando diversi layout e formati di rappresentazione, dipendentemente dall'uso fatto.

Parte II

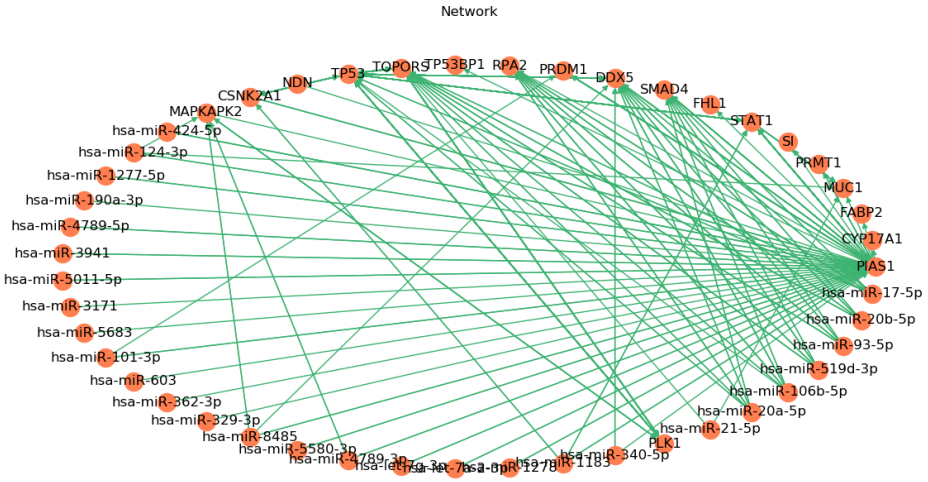
Analisi di reti di co-espressione genica

La presente sezione ha l'obiettivo di spiegare gli algoritmi pensati e sviluppati con il fine di semplificare l'osservazione dei grafi di studio.
La caratteristica peculiare che accomuna questi codici sono la loro versatilità, in quanto non sono stati pensati solo per l'analisi di geni coinvolti nel disturbo dello spettro autistico, ma possono essere utilizzati in altri contesti di studio.
La stesura dei codici è avvenuta in un ambiente di sviluppo con linguaggio *Python*, in previsione di poter essere trasferiti in una locazione fruibile a terzi in futuro, come può essere un sito web direttamente collegato al *The RING database*.
Per questo motivo la grafica non rispecchia quella del sito di *The RING database*, ma è compatibile con gli stili della libreria *matplotlib*.

Un esempio della diversa grafica su cui si è basato lo studio e che apparirà nelle pagine seguenti è visibile in Figura 4.5.



(a) *Layout prodotto dal The RING database [12]*



(b) *Layout circolare prodotto con matplotlib*

Figura 4.5: *Grafi dei transcription factors PIAS1 e TOPORS, tutte le loro interazioni e connessioni*

Capitolo 5

Divisione per proprietà

Direzione

Ogni arco possiede un'informazione riguardante la *direzione*, che rappresenta la generica azione compiuta dall'arco stesso.
Tra i diversi simboli si possono trovare:

- —>
  Relazione di signalling d'attivazione, il nodo sorgente quando espresso attiva il nodo target. Se l'attivazione viene compiuta da un transcription factor prima di questo simbolo si troverà la lettera 't', ottenendo così **t**—>.

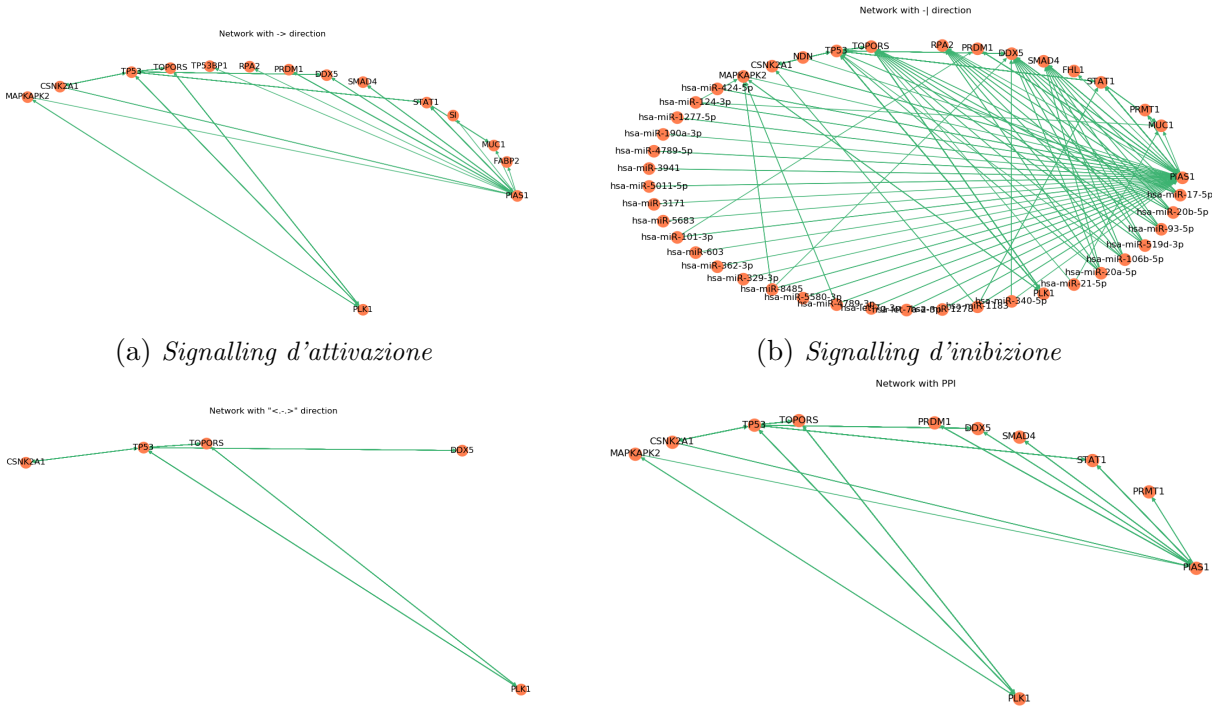
- —|
  Relazione di signalling d'inibizione, il nodo sorgente quando espresso inibisce il nodo target. Molto spesso si trova come **m**—|, dove la lettera 'm' rappresenta l'inibizione da parte di un miRNA, che per definizione è un frammento di RNA non codificante il quale ha un'azione inibitrice delle zone a cui si lega. Talvolta si trova anche il simbolo **t**—| che rappresenta un'inibizione causata da un transcription factor.

- <.—.>
  Co-attivazione, il nodo sorgente e quello target si attivano a vicenda dando luogo a un complesso d'attivazione.

- —
  Protein protein interaction, PPI; questa relazione è a livello di proteoma.

Questi tre simboli rappresentanti l'azione dell'arco sono stati scelti per fare una prima classificazione delle entità appartenenti al network studiato.
La funzione che permette di fare ciò prende il nome di *division_direction.py* e consiste nell'analizzare ogni arco del grafo estraendo solo il campo *direction*. Gli archi vengono successivamente divisi in tre categorie: attivazione, inibizione e complesso d'attivazione.

La presente funzione è stata pensata in modo da permettere all'utente di poter accedere a un confronto rapido e semplice delle relazioni a livello di genoma o di proteoma tra le entità di studio, come si può vedere in Figura 5.1.



(a) *Signalling d'attivazione*

(b) *Signalling d'inibizione*

(c) *Co-attivazione*

(d) *Protein Protein Interaction*

Figura 5.1: *Output grafico della funzione division_direction.py con l'intento di suddividere le entità a seconda del simbolo di direzione*

Azione

Un'altra informazione che viene riportata sugli archi è la *action*, o meglio la relazione chimico-biologica alla base dell'arco.
Basando l'idea su queste differenze di informazioni è nato il metodo di semplificazione che si basa appunto sulla separazione delle azioni, contenuto nella funzione *division_action.py*.
Le azioni, viste come informazioni del grafo, sono contenute negli archi che collegano i nodi, perché esprimono la loro relazione.
Tra questi rapporti possiamo trovare:

- fosforilazione
- catalisi
- regolazione con miRNA
- repressione
- espressione regolata
- associazione
- associazione fisica
- predizione
- interazione diretta

Nel presente elenco si può notare come gli ultimo 4 elementi siano molto vaghi, non danno informazioni ben precise su quello che accade biologicamente tra le due entità, ma nello stesso tempo affermano che è presente una correlazione tra di essi e questa è un'indicazione già di per sè.
É stato scelto di privilegiare le interazioni quali fosforilazione, catalisi, regolazione con miRNA, repressione ed espressione regolata, andando a produrre per l'appunto dei grafi a sè stanti riprodotti poi a video e anche dei pandas dataframes facilmente convertibili in file di testo.
Gli altri elementi tuttavia non sono stati del tutto ignorati ed eliminati, ma sono stati racchiusi in un grafo a parte chiamato *others* e inoltre sono stati salvati in un file *.csv* contenente tutte le loro proprietà.
Il file di testo permette pertanto una consultazione per un'analisi postuma, con l'idea di modificare in futuro l'informazione vaga con un'etichetta più esplicativa.

Tutto ciò è stato sviluppato con un semplice controllo del nome dell'azione contenuto negli archi, come nella funzione precedente, e possibile solo grazie alla standardizzazione delle nomenclature all'interno del *The RING database* ed è stata pensata per permettere una consultazione rapida delle funzioni biochimiche agenti tra le varie entità, Figura 5.2.

(a) *Fosforilazione*

(b) *Catalisi*

(c) *Espressione regolata*

(d) *Regolazione con miRNA*

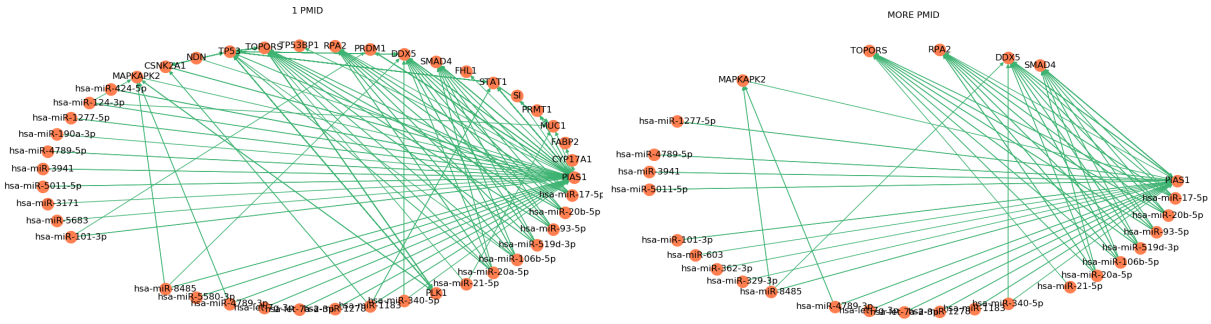
(e) *Repressione*

(f) *Altri*

Figura 5.2: *Output grafico della funzione division_action.py con l'intento di suddividere le entità a seconda della reazione biochimica*

Numero di pubblicazioni

Anche questa strategia di semplificazione, come le due precedenti, non prevede basi di utilizzo di variabili informatiche.
Lo sviluppo del filtro, contenuto nella funzione *division_pmid.py*, si fonda su una semplice analisi del numero di pubblicazioni di ogni cammino. L'attributo PMID, acronimo di *PubMed IDentifier*, è contenuto nei file dei nodi e degli archi, per permettere all'utente di poter avere un'immediata consultazione della pubblicazione che ha portato alla nascita del dato presente. Questa etichetta è stata dunque soggetto della discriminazione tra due tipi di grafi: il primo possedente al massimo una sola pubblicazione, il secondo avente due o più.
Il numero di pubblicazioni è anche una sorta di controversia, in quanto alcuni esponenti della ricerca scientifica tendono ad avere la smania di riconoscimenti e perciò arrivano a produrre un numero elevato di papers. Articoli che molto spesso non aumentano l'informazione già esistente su quell'argomento, ma che lo ripetono e lo solidificano ancora di più.
Il presente modo di approcciarsi alla ricerca crea una sorta di stallo, bloccando così l'innovazione e facendo crescere l'idea che più pubblicazioni si hanno più si è un ricercatore di fama, portando così giovani scienziati a ricercare la strada più facile.
Per questi motivi sono nati i due grafi e due pandas dataframes separati: il primo, con una sola pubblicazione, utile per invogliare lo studio riguardante quelle relazioni; il secondo, con più pubblicazioni, adatto per ricevere conferme su osservazioni ottenute, vedi Figura 5.3



(a) *Singola pubblicazione* (b) *Più pubblicazioni*

Figura 5.3: *Output grafico della funzione division_pmid.py con l'intento di suddividere le entità a seconda del numero di pubblicazioni*

Capitolo 6

Complessi proteici

Cliques

Finite le divisioni sui vari campi presenti all'interno degli archi, ci si è spinti verso un'analisi basata su variabili informatiche.
L'idea si basa sull'estrazione delle *cliques*, cioè gruppi in cui i suoi elementi sono tutti interconnessi tra di loro, ed è contenuta nella funzione *division_cliques.py*.
Geometricamente parlando si può pensare alla forma basilare di un segmento, i cui estremi sono rappresentati da due nodi; oppure da un triangolo, dove ciascun vertice è collegato agli altri due da un semplice arco; fino a spingersi verso figure più complesse, basta che ogni elemento sia connesso con tutti gli altri.
Biologicamente parlando, si possono vedere questi gruppi di elementi tutti connessi tra di loro come dei complessi proteici ad esempio. Geni le cui proteine lavorano e si influenzano l'una con l'altra per dare origine a fenomeni biologici.
Da questa riflessione è nata la funzione di divisione in base alle cliques, la quale prevede la costruzione di un grafo temporaneo con solo le PPI, facendo uno studio solo a livello di proteoma. L'analisi viene fatta ricercando le cliques presenti all'interno di questa rete, andando però a scartare tutti quei complessi composti da solo due elementi connessi tra loro. La scelta di eliminare queste cliques si basa sul fatto che il grafo non avrebbe subito granché mutazione, andando quindi a prediligere una maggiore semplificazione di questo. Ogni clique composta da più di due elementi viene memorizzata, e in seguito viene mostrato all'utente un sottografo contenente tutte le cliques. Questo grafo però non le rappresenta in modo diviso, ma anzi mostra le interconnessioni tra i vari gruppi, permettendo dunque una riflessione sulle varie interazioni tra i diversi complessi. Nel caso della Figura 6.1 però è stato trovato un solo complesso.

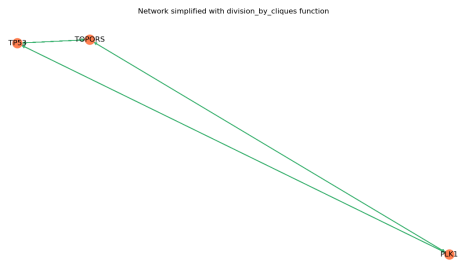


Figura 6.1: *Output grafico della funzione division_cliques.py con i complessi proteici*

Coefficiente di clustering

L'identificazione e l'estrazione di strutture costituenti grafi completi, dette cliques, che indicano complessi proteici o strutture ad alta interazione, potrebbe rappresentare un ottimo strumento per condurre uno studio più approfondito e selettivo di tali sotto-sistemi, permettendo così di rispondere a diversi interrogativi biologici riguardanti il fenomeno analizzato. Un'altra strategia di processing del grafo, mirata a tale scopo, è stata elaborata basandosi sul calcolo iterativo di un parametro proprio di ogni nodo: il Coefficiente di Clustering Locale.

Il Coefficiente di Clustering rappresenta la tendenza del grafo ad essere diviso in gruppi, detti cluster. Esso può essere Globale o Locale. Quello utilizzato in questo algoritmo è il secondo, che viene calcolato per ogni nodo e restituisce il rapporto tra il numero di relazioni esistenti fra i vicini del nodo e il numero massimo di archi che potrebbero esistere tra gli stessi, rappresentando quindi la tendenza dei vicini di un nodo, congiuntamente al nodo stesso che risulta collegato ad essi, a costituire un sottografo completo [2].

Questo parametro risulta, dunque, essere un elemento valido per il calcolo iterativo di cricche presenti nel grafo. L'algoritmo proposto, infatti, elabora il coefficiente di clustering di ogni nodo presente sul grafo iniziale, mostrato in Figura 6.2, procedendo successivamente all'identificazione dei nodi aventi valore massimo e, dunque, costituenti, insieme ai propri vicini, sottografi completi.

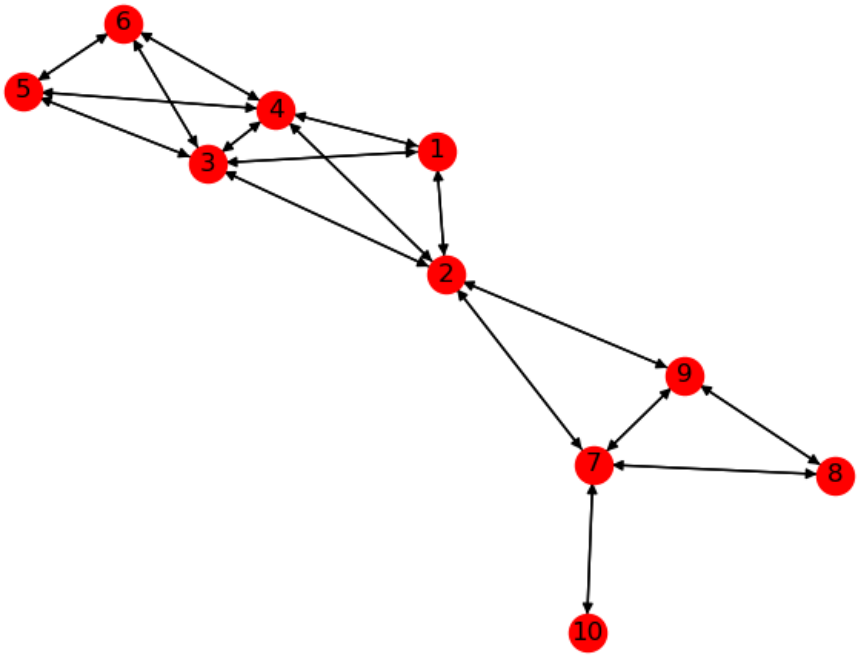
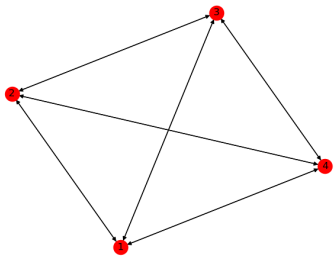
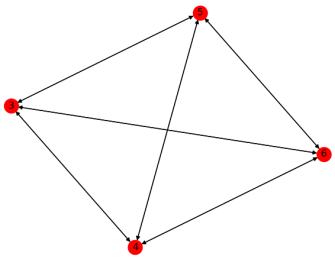


Figura 6.2: *Rete di partenza da elaborare con l'algoritmo clust_subgs.py*

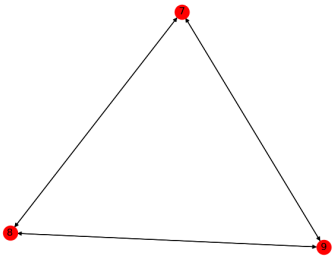
Per ognuno di questi nodi, viene quindi creata una lista comprendente esso e i nodi adiacenti e ne viene estratto il sottografo, come mostrato nella Figura 6.3, salvato successivamente in un'altra lista, contenente tutti i sottografi elaborati, a meno che esso non sia stato già identificato processando un altro nodo.



(a) *Ottenuto dal nodo 1*



(b) *Ottenuto dai nodi 6 e 5*



(c) *Ottenuto dal nodo 8*

Figura 6.3: *Cliques estratte, mediante l'algoritmo clust_subgs.py, trovando i nodi con coefficiente di clustering pari ad 1*

In seguito, tutti i nodi processati, vengono eliminati dal grafo iniziale, generando così un nuovo grafo da poter elaborare nuovamente nelle iterazioni successive, come mostrato in Figura 6.4.

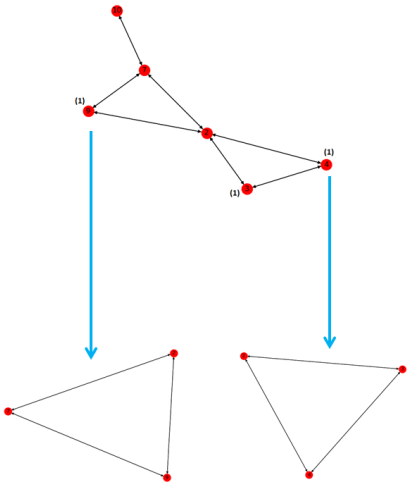


Figura 6.4: *Seconda iterazione effettuata dall'algoritmo clust_subgs.py, con relativa estrazione delle cliques identificate.*

L'algoritmo procede iterativamente, finché esiste almeno un nodo appartenente al grafo da processare e avente, come coefficiente di clustering, valore 1.
In output vengono dunque restituiti tutti i sottografi delle cliques estratte e il grafo finale, mostrato in Figura 6.5, impossibile da processare ulteriormente, ma rappresentante le interconnessioni tra le varie strutture.

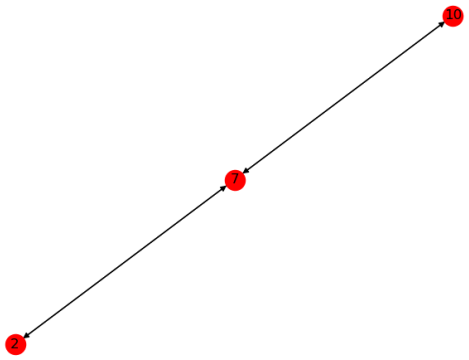


Figura 6.5: *Grafo non ulteriormente riducibile dall'algoritmo clust_subgs.py, raffigurante le connessioni residue.*

L'iteratività di questo algoritmo permette di studiare, in maniera dettagliata e ravvicinata, tutti i diversi complessi estraibili, nonché alcune porzioni di essi, mantenendo però una mappatura delle connessioni, costituita dal grafo finale non ulteriormente elaborato.

Capitolo 7

Esplorazione path

L'algoritmo proposto esegue varie tipologie di esplorazione di percorsi, detti Paths.
Un path è un grafo semplice nel quale i nodi sono ordinati in modo tale da avere nodi adiacenti solo se essi sono consecutivi all'interno della lista di appartenenza [3].
I cammini identificati possono essere appartenenti a 2 classi, identificate dalla libreria Networkx:

- **Simple paths**: sono percorsi nel quale non esistono ripetizioni di nodi.
- **Shortest paths**: percorsi per il quale la somma dei pesi dovuti ad ogni arco è minimizzata.

Altre definizioni utili per l'analisi di queste routine sono le seguenti:

- **Centro**: è un sottografo composto dai vertici che restituiscono massima eccentricità, dove il massimo di questa è uguagliata dal diametro del grafo.
- **Degree Centrality**: è rappresentato dal numero di interconnessioni uscenti o entranti da ogni nodo.
- **Hub**: rappresentano i nodi con Degree Centrality più altri, quindi ricchi di interconnessioni incidenti ad esso.
- **Foglie**: rappresentano i nodi estremi del grafo.

In molte ricerche di ambito biologico, può essere interessante ricercare tutti i percorsi esistenti tra una determinata sorgente e un target, al fine di studiare le relazioni esistenti tra essi. Ciò infatti potrebbe consentire di determinare, ad esempio, il modo in cui l'azione di gene o di una qualsiasi entità biologica influenza l'azione di un'altra entità, mediante cammini regolatori. Anche l'esplorazione automatizzata di percorsi riguardanti componenti di grande interesse, come Hub o nodi terminali, potrebbe essere di grande supporto nello studio dei network biologici.

È da queste considerazioni che nasce l'algoritmo di esplorazione dei percorsi, suddiviso in 4 diverse funzioni:

- Source to Target (through a Node)
- Center(s) to Leaves/Leaves to Center(s)
- Hub(s) to Leaves/Leaves to Hub(s)
- Leaves to Leaves

Source to Target (through a Node)

La prima funzione del modulo permette di investigare tutti i percorsi esistenti tra una sorgente ed una destinazione, in diversi modi, prendendo in ingresso la rete, il nodo di origine e la destinazione.
Inizialmente, essa va a verificare l'esistenza di un simple path tra la sorgente e il target; nel caso in cui questa condizione fosse verificata, si procede con l'elaborazione di tutti i percorsi semplici, partendo dal più breve, salvando i percorsi congruenti in una lista apposita. Nel caso in cui non esistessero, invece, percorsi semplici, ma esistesse uno o più percorsi brevi, la routine procede all'elaborazione di tutti gli shortest paths, salvando i percorsi congruenti in una lista apposita. In Figura 7.1 è mostrato il grafo di partenza, affiancato da tutti i percorsi identificati tra il nodo source e quello target.

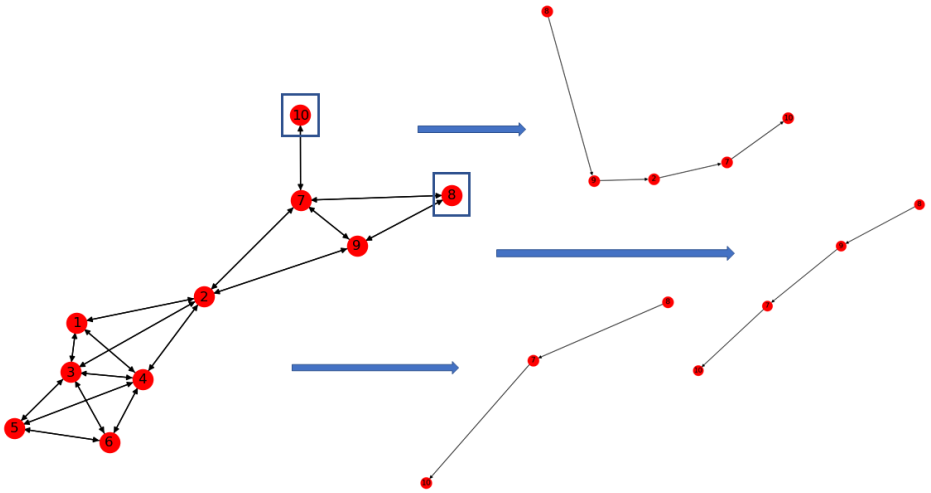


Figura 7.1: *Grafo di partenza e percorsi identificati con la funzione Source_Target (ExpPath.py), scegliendo, rispettivamente, i nodi 8 e 10 come source e target*

Questa funzione presenta inoltre 2 parametri addizionali: la lunghezza massima ed un nodo intermedio, o di transito.

Il parametro di lunghezza massima permette di salvare nelle apposite liste, quando modificate, solo i percorsi presentanti un numero di nodi minore o uguale al valore inserito, riducendo il numero di percorsi calcolati. Il valore di default impostato è pari a 1000. In Figura 7.2 è mostrato il grafo di partenza, affiancato dai percorsi identificati tra i 2 nodi scelti che rispettano la condizione sulla lunghezza massima pari a 5. Ovviamente la presenza di questo valore potrebbe far sì che le liste siano restituite come vuote, nel caso in cui non esistesse nessun percorso rispettante tale condizione, ecco perché è stato scelto un valore di default abbastanza elevato.

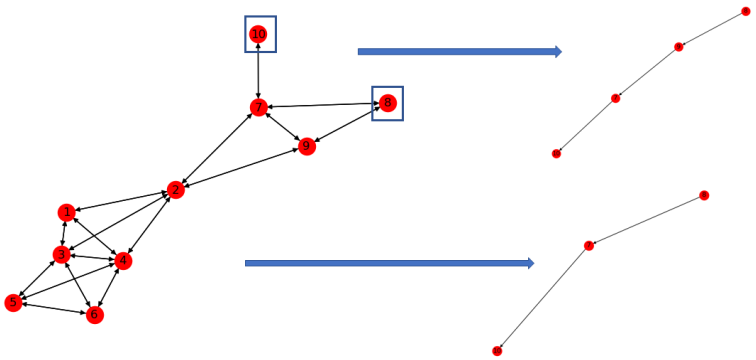


Figura 7.2: *Grafo di partenza e percorsi identificati con la funzione Source_Target (ExpPath.py), scegliendo, rispettivamente, i nodi 8 e 10 come source e target e lunghezza massima pari a 4*

Il parametro di "nodo di transito", permette invece di generare un'altra lista, contenente, però, i percorsi tra source e target e transitanti per questo ulteriore nodo. In Figura 7.3 è mostrato il grafo di partenza, affiancato dai percorsi identificati tra i nodi e transitanti per il nodo scelto.

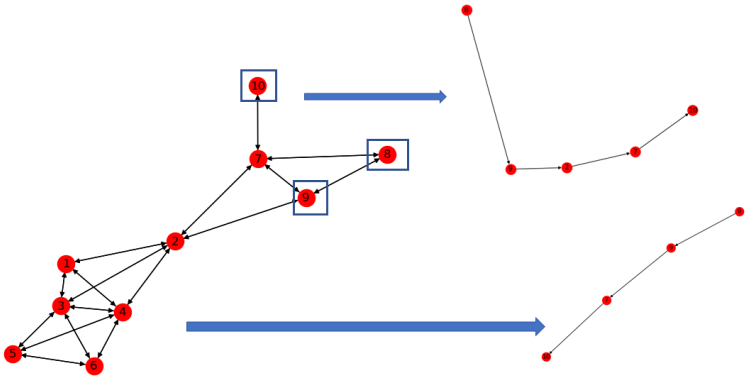


Figura 7.3: *Grafo di partenza e percorsi identificati con la funzione Source_Target (ExpPath.py), scegliendo, rispettivamente, i nodi 8 e 10 come source e target e il nodo 9 come transito*

I percorsi "con transito" vengono generati contemporaneamente a quelli tra source e target, quindi la routine implementa un ulteriore controllo, nel caso in cui non esistessero percorsi semplici transitanti per quel nodo, pur esistendo tra la sorgente e la destinazione, verificando l'eventuale esistenza di percorsi brevi. La contemporaneità della generazione dei 2 tipi di percorsi, fa si che anche questi possano essere condizionati dal parametro di lunghezza massima. In Figura 7.4 è mostrato il grafo di partenza, affiancato dai percorsi identificati, presentanti il transito e rispettanti la condizione sulla lunghezza massima, pari a 4.

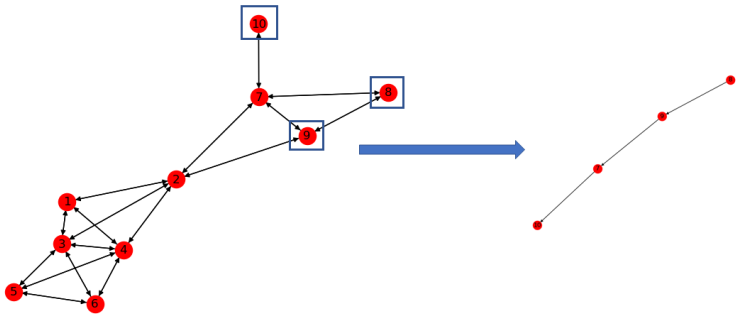


Figura 7.4: *Grafo di partenza e percorsi identificati con la funzione Source_Target (ExpPath.py), scegliendo, rispettivamente, i nodi 8 e 10 come source e target,il nodo 9 come transito e lunghezza massima pari a 4*

Center(s) to Leaves/Leaves to Center(s)

La seconda funzione del modulo permette di identificare tutti i percorsi esistenti tra il Centro del grafo e i terminali dello stesso, in ambedue le direzioni, prendendo in ingresso la rete di partenza. Dopo aver verificato che il grafo sia strettamente connesso e averne calcolato il/i Centro/i, si procede con la distinzione delle foglie in due categorie: le foglie senza predecessori, che rappresentano le foglie iniziali, e quelle senza successori, che rappresentano le foglie terminali. I percorsi, infatti, vengono calcolati, come detto, in 2 direzioni: una prende iterativamente come source le foglie iniziali e come target il/i Centro/i, l'altra prende iterativamente come target le foglie terminali e come source il/i Centro/i. I percorsi vengono calcolati in maniera analoga a quanto visto precedentemente, a meno della condizione sulla lunghezza massima, che sembrava troppo variabile e, dunque, poco prevedibile. In uscita, vengono restituite le due liste, con i percorsi dal/i Centro/i alle foglie e viceversa.

Hub(s) to Leaves/Leaves to Hub(s)

La terza funzione del modulo permette di identificare tutti i percorsi esistenti tra uno o più Hub del grafo e i terminali dello stesso, prendendo in ingresso la rete. Anche questa funzione, come la precedente, computa i percorsi in ambedue le direzioni, distinguendo le foglie in iniziali e terminali. Inizialmente la viene calcolata la Degree Centrality di ogni nodo nella rete, identificandone il valore massimo tra i dati ottenuti e mappandolo a tutti i nodi riportanti lo stesso valore. Successivamente, vengono calcolati, sempre in maniera iterativa, tutti i percorsi andanti dal/dagli Hub alle foglie terminali e viceversa, considerando le foglie iniziali. Anche questi percorsi vengono calcolati facendo il doppio controllo sull'esistenza di simple e shortest paths e facendo a meno della condizione sulla lunghezza massima. In uscita vengono restituite le due liste, con i percorsi dal/dagli Hub alle foglie e viceversa.

Leaves to Leaves

La quarte funzione del modulo permette di identificare tutti i percorsi esistenti tra le varie foglie, prendendo in ingresso la rete. Ovviamente, anche in questo caso le foglie che fungono da source saranno quelle iniziali, mentre quelle che fungono da target saranno le terminali. Iterativamente vengono calcolati tutti i percorsi richiesti, mantenendo il controllo sull'esistenza dei simple e degli shortest paths, e aggiunti ad una singola lista, fornita in Output. Anche in questo caso, viene fatto a meno della lunghezza massima, che rappresenterebbe nuovamente un parametro instabile.

Capitolo 8

Componente fortemente connessa

Un ulteriore algoritmo sviluppato si basa sull'identificazione di una o più componenti fortemente connesse, quando esistenti, all'interno del grafo.
Teoricamente, un grafo fortemente connesso è un grafo nel quale per ogni coppia ordinata di nodi esiste un percorso orientato tra questi. Le componenti fortemente connesse di un grafo risultano, quindi, esserne i sottografi massimali fortemente connessi, per i quali esiste un percorso orientato congiungente ogni coppia di nodi ordinata. [3]
Questa strategia, ovviamente, richiedendo l'esistenza di un cammino orientato, è applicabile solo a grafi di tipo direzionato, che, come detto nel Capitolo 4, possono rappresentare solo alcuni tipi di reti biologiche, come quelle metaboliche o regolatorie. Sebbene essa sia un'implicazione di difficile esistenza in campo biologico, l'identificazione di una componente fortemente connessa permetterebbe una notevole semplificazione dello studio del sistema, consentendo l'estrazione e l'analisi di essa, in maniera indipendente dagli altri elementi presenti nella rete di partenza, e il suo "collasso" in un nodo singolo, che sostituisce così l'intera componente nel grafo primordiale.
É proprio questo che si prefigge di fare la strategia di semplificazione proposta.

Partendo da un grafo iniziale, mostrato in Figura 8.1, si estraggono tutte le componenti fortemente connesse presenti in esso, aggiungendole in una lista dedicata, a meno che esse non abbiano un numero di nodi pari o inferiore ad 1.

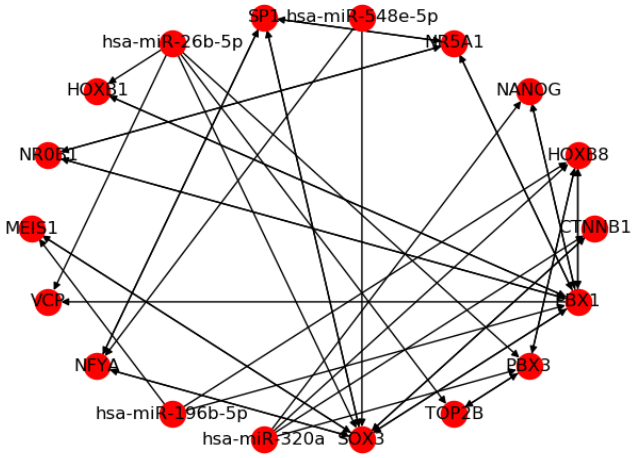


Figura 8.1: *Rete iniziale da elaborare con algoritmo Stron_Conn_sub.py*

Si collassano poi tutte le componenti connesse identificate in un singolo nodo, semplificando il grafo iniziale e generandone uno detto Condensato, mostrato in Figura 8.2. Giacchè il grafo condensato subisce una modifica dei nomi e degli archi in seguito all'estrazione, è stata implementata una strategia di "remapping" di tutti i nomi dei nodi rappresentanti le diverse componenti fortemente connesse.

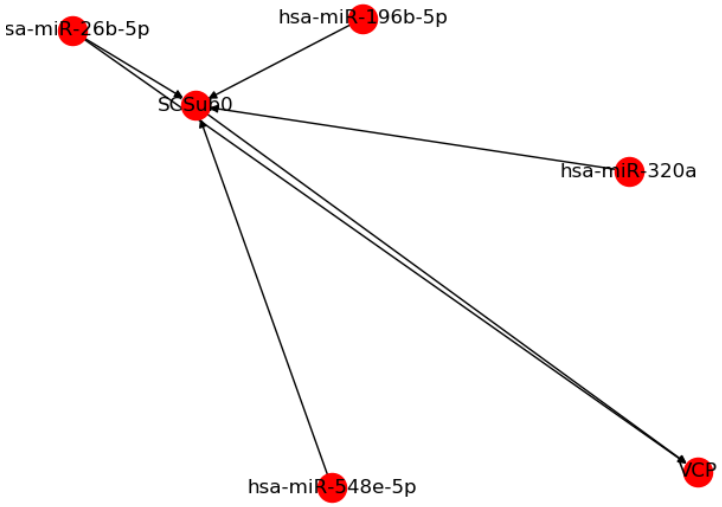


Figura 8.2: *Versione condensata e remapped del grafo iniziale ottenuta con algoritmo Strong_ Conn_ sub.py*

In output all'algoritmo, vengono restituiti separatamente il grafo condensato e le componenti fortemente connesse, mostrate in Figura 8.3.

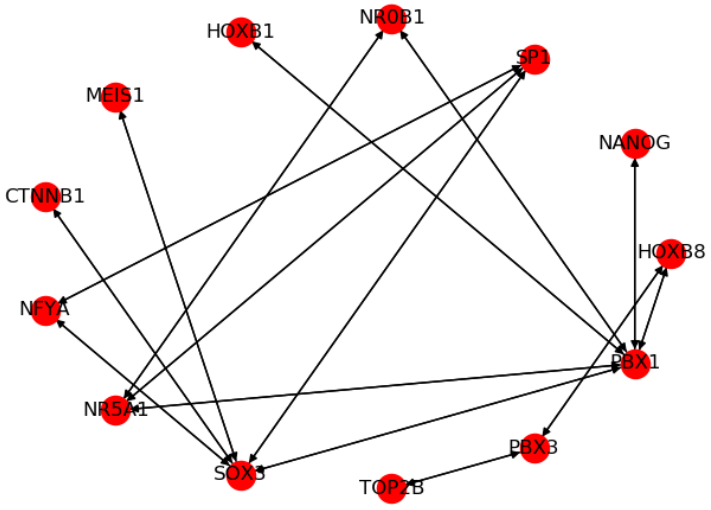


Figura 8.3: *Componente fortemente connessa SCSub0 estratta con algoritmo Strong_ Conn_ sub.py*

Capitolo 9

Betweenness

Le reti bioinformatiche sono sistemi molto complessi e imprevedibili, questa difficoltà nello studio fa sì che si debba ricorrere all'analisi di diversi coefficienti che aiutano a decifrare il comportamento del grafo in modo del tutto automatico.
Sono stai studiati 4 coefficienti diversi:

- *Degree Centrality*
- *In Degree Centrality*
- *Out Degree Centrality*
- *Betweenness*

Il grado di centralità è di solito uno tra i valori più utilizzati nello studio delle reti. Il suo valore dipende esclusivamente dal numero di archi, o edges, che dipartono e arrivano al nodo stesso. Mentre la degree centrality permette uno studio senza considerare le varie direzionalità degli archi, l'in degree centrality tiene conto delle entità entranti, al contrario l'out degree centrality di quelle uscenti. Questi particolari valori però per essere calcolati richiedono un grafo molto connesso, condizione che non è sempre verificata all'interno delle reti bioinformatiche. Un grafo, o un suo sottoinsieme, è connesso quando tutti i suoi elementi sono collegati tra di loro attraverso degli archi; ma questa condizione non avviene tutte le volte all'interno di interconnessioni biologiche, in quanto dipendono dal loro rapporto ma anche da una eventuale non scoperta.
Per questo motivo, dopo uno studio non esaustivo anche di differenti valori di centralità, ho deciso di scartare questo tipo di analisi, e concentrarmi interamente sulla betweenness.
La betweenness tiene conto dei diversi cammini, cioè se per spostarsi da un nodo all'altro si tende a passare più volte per un particolare nodo, quello avrà un valore di betweenness elevata in quanto si trova in una posizione strategica importante.
I nodi con alta betweenness sono chiamati *bottlenecks* e si trovano di solito tra nodi con molti archi entranti e uscenti, rappresentando un ponte di comunicazione tra questi nodi di elevata importanza [5].
Biologicamente parlando, un nodo con alta betweenness può essere interpretato come una proteina la quale porta al blocco di un processo se non attiva; si può dunque usare questo coefficiente per conferire un valore di essenzialità.

Analisi bottlenecks e nodo ponte

La funzione *one_link.py* riceve in ingresso il grafo originario e come prima cosa calcola la betweenness centrality, la quale per semplicità verrà chiamata *betwenness*.
Si è poi proceduto con il mettere in ordine i valori trovati in modo tale da analizzare per primi i nodi con un valore alto di betwenness.
Ogni singola betwenness è stata analizzata, andando a procedere verso i passi successivi con solo i nodi aventi un valore diverso da zero. Lo zero è stato scelto per semplicità, ma non si scarta l'ipotesi di poter mettere un valore più alto, e dunque poter così semplificare molto di più quello che sarà il grafo finale; tuttavia non è stato scelto un valore diverso da zero per via di una idea di semplificazione un po' troppo forzata che dunque avrebbe portato a un risultato troppo approssimato.
Di ogni nodo con una betwenness diversa da zero, si va poi a salvare il suo intorno, andando a ricercare negli archi direzionati del grafo originario la sua presenza sia con la freccia entrante sia con quella uscente.
Questa piccola unione simil fiore dandelion è stata chiamata *foresta*, dove il nodo centrale è il ramo, e gli altri nodi collegati a lui sono le foglie.
Dopo aver salvato tutte le foreste, si è proceduto con l'analisi di queste ultime a coppie di due, ricercando la loro intersezione.
Le foglie con la funzione di ponte tra due foreste sono state perciò salvate. Ogni intersezione è stata soggetta a un controllo di presenza o meno di oggetti, e poi in caso di presenza di nodi comunicanti si è proceduto con il salvataggio singolo per poi dare vita al nuovo grafo.
La rete risultante avrà dunque una serie di nodi con una betweenness alta, e dunque con una cerca importanza nella loro ricorrenza in varie attività, e in più qualche nodo che permette a loro di comunicare, vedi Figura 9.1.
Una possibile modifica è la ricerca del nodo collegante diversa da miRNA, quindi un ulteriore filtro che aumenta la selettività dei nodi che andranno a far parte del grafo finale.

Dal punto di vista biologico lo si può interpretare come una ricerca di elementi comuni tra due entità importanti, e una possibile visione immediata di ripercussioni che si potrebbero avere se uno di questi tre elementi subisse una mutazione.

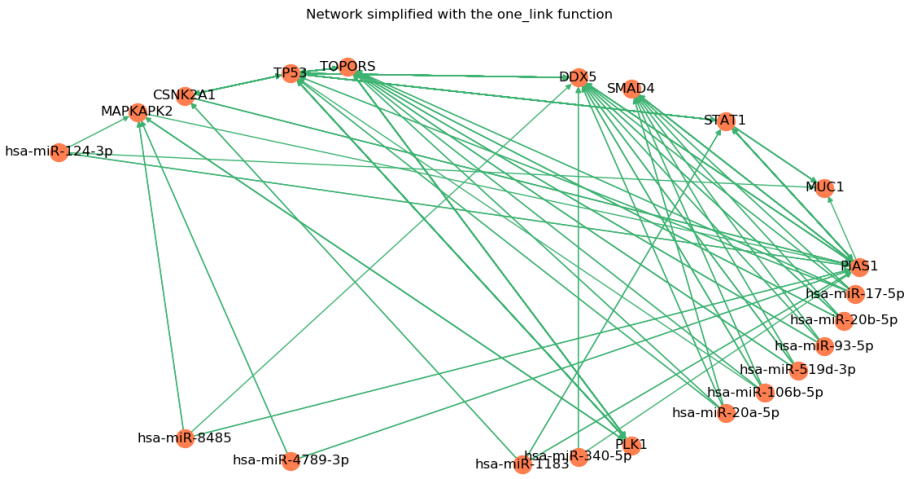


Figura 9.1: *Output grafico della funzione one_link.py con ritenzione di elementi bottlenecks e nodi ponte*

Analisi bottlenecks senza miRNA e nodo ponte

I microRNA o miRNA sono frammenti non codificanti di RNA maturo, di circa 22 basi nucleotidiche, che influenzano i geni a cui si legano, inibendo l'espressione genica del sito complementare.
Queste molecole talvolta non sono considerate da molti ricercatori, pertanto alcuni geni tendono ad avere numerosi miRNA associati, mentre altri molti meno.
Per evitare questa influenza è stato introdotto un filtro apposito.

La betweenness, come detto precedentemente, dipende anche da quanti nodi sono connessi al nostro elemento d'interesse. Talvolta capita che molti miRNA si trovino ad essere tutti interconnessi ad un singolo elemento, facendolo apparire di grande interesse per via della sua betweenness alta, ma in realtà con pochi altri elementi biologici connessi. Per questo motivo, prima di riproporre la funzione precedente, si è inizializzato un grafo temporaneo che non tiene conto dei miRNA.
Oltre ad eliminare l'influenza dei miRNA si è rimosso l'effetto sul valore di betweenness da parte di archi ripetuti, condizione che si verifica per il fatto che si tratta di un *Multi-DiGraph*. Per fare in modo che non ci sia nessuna influenza sull'analisi dei bottlenecks, che non sia per via di connessioni con geni o transcription factors, il grafo temporaneo senza miRNA è stato inizializzato a *DiGraph*.
Non è stato scelto di mettere a priori questo dato filtro per mantenere così la facoltà di scelta all'utente e nello stesso tempo dare la possibilità anche di confrontare il risultato della funzione *one_link.py* con quello dell'attuale funzione *one_link_no_mirna.py*.
Il procedimento è il medesimo del precedente, con uno studio dei nodi con valore diverso da zero e poi la memorizzazione dei nodi centrali e del nodo collegante.
Anche qui è possibile la modifica del valore di soglia della betweenness e inoltre l'eventualità di non considerare i miRNA negli elementi di ponte tra le due entità di interesse.
In Figura 9.2 è possibile notare come a differenza della Figura 9.1 non sia presente l'elemento SMAD4 (settimo elemento iniziando a contare dal miRNA in alto a sinistra).

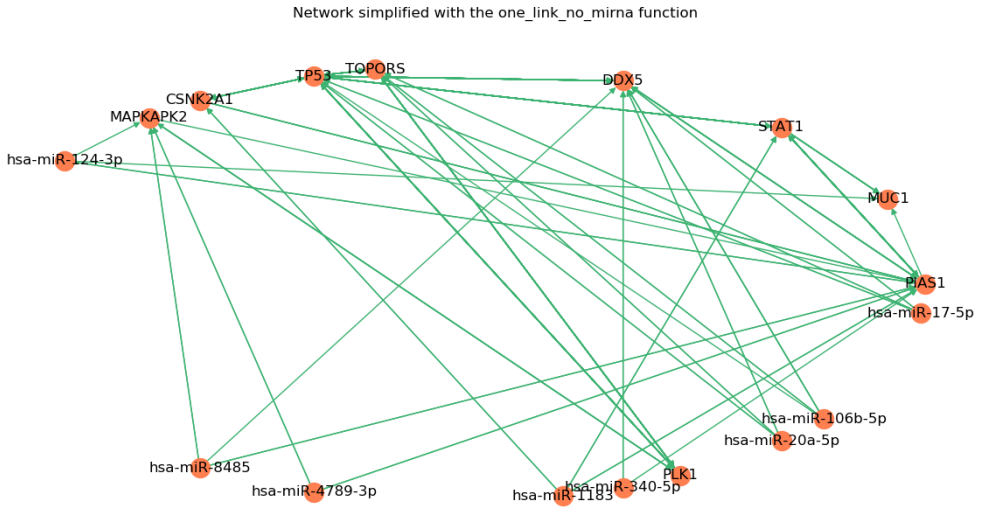


Figura 9.2: *Output grafico della funzione one_link_no_mirna.py.py con ritenzione di elementi bottlenecks (trovati senza miRNA e senza ripetizioni di archi) e nodi ponte*

Parte III

Simulatori di reti di co-espressione genica coinvolte nell'autismo

Ci sono numerose tecniche di studio che riguardano una rete bioinformatica: una è quella vista nel Parte II e prevede una semplificazione del grafo in modo da estrapolare informazioni velocemente e in modo più semplice per l'utente, un'altra tecnica può prevedere la costruzione di una rete di simulazione.
Questo studio si basa sull'analisi dell'espressione dei geni all'interno del network, che possono essere: espressi, repressi e una via di mezzo.

Come primo passaggio si può pensare al non considerare le vie di mezzo, cioè vedere la rete come un'insieme di elementi che vengono espressi o repressi. Il passaggio successivo è trasformare questo problema biologico in uno informatico, e quindi vedere un gene espresso come una variabile attiva a cui si può far corrispondere il valore pari a 1, e un gene represso come una variabile spenta con costante pari a 0.
In seguito si analizzano i vari archi congiungenti i nodi e si osserva se questi sono di *signalling d'attivazione*, *signalling d'inibizione* o riguardano il proteoma e sono *PPI*.
Avendo due possibili valori del nodo e tre eventuali tipi di arco, si hanno sei combinazioni:

- *gene espresso e arco di attivazione*
  Il risultato sarà l'attivazione del nodo target.

- *gene represso e arco di attivazione*
  Il risultato sarà l'inattivazione del nodo target.

- *gene espresso e arco di inattivazione*
  Il risultato sarà l'inibizione del nodo target.

- *gene represso e arco di inattivazione*
  Il risultato sarà l'attivazione del nodo target.

- *gene espresso e arco PPI*
  Il risultato sarà indipendente da quest'arco, non ci da informazioni in più.

- *gene represso e arco PPI*
  Come il precedente.

Nei paragrafi seguenti verranno spiegate diverse idee di implementazione di un programma che simula la propagazione dell'espressione.
Il primo simulatore essendo un prototipo si basa su dati inventati, pertanto sono stati creati file appositi con pochi campi.
Per quanto riguarda il simulatore definitivo, ci sono state difficoltà computazionali e sono stati usati file diversi dai file *edges.csv e nodes.csv*; si è infatti partiti da file pickles aventi due campi oltre al nome: *direction* e *int_count* (rappresentante quante volte quell'arco è ripetuto).

Capitolo 10

Prototipo di un simulatore per confronto con un golden standard

L'idea su cui si è fondata la costruzione di questo simulatore è il confronto del network con un golden standard, o meglio, di un file contenente tutte le espressioni dei nostri geni. I file utilizzati per lo sviluppo sono due: uno che simula l'output *edges.csv* del The RING Database e quindi che riprende la standardizzazione degli elementi, con campi ridotti a tre per semplicità; l'altro riporta invece i nomi dei nodi e tra le altre caratteristiche simili al file *nodes.csv* aggiunge l'informazione sull'espressione, quindi un valore che corrisponde allo stato di espressione. Tutto ciò è visibile in Figura 10.1.

source_name	target_name	direction
A	B	->
B	C	-|
B	F	->
B	E	->

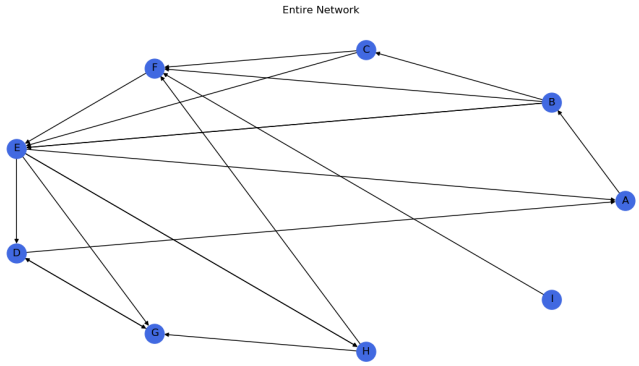
(a) *File per costruzione del grafo che simula edges.csv*

node_name	status
A	*1*
B	*1*
C	*0*
D	*1*

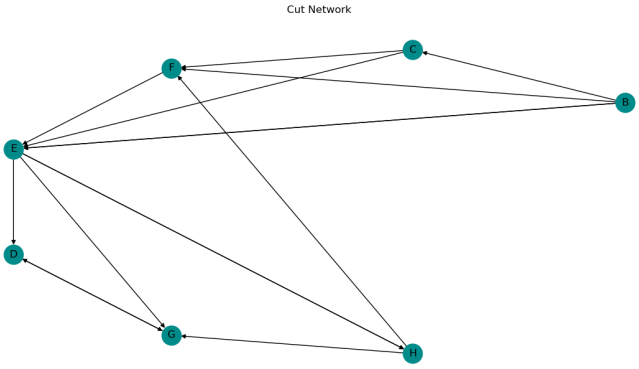
(b) *File 'Golden Standard' contenente le espressioni dei nodi*

Figura 10.1: *Estratti dei file creati per sviluppare il simulatore per confronto con Golden Standard*

L'azione preliminare che viene fatta dopo la lettura dei due file di nodi e archi, è il taglio dell'intero grafo. La semplificazione viene fatta in base a quale nodo viene reputato sorgente e su quale nodo vogliamo studiare gli effetti dell'espressione del nodo sorgente. L'utente attraverso command window identifica questi due elementi e il software procede ricercando tutti i percorsi semplici che collegano le due entità. Questo viene fatto sia per diminuire i tempi computazionali, sia per avere una certa metrica di inizio di propagazione dei risultati, senza la quale ogni punto dovrebbe essere di partenza e/o arrivo, andando a complicare notevolmente il risultato.
Bisogna però affermare che questa semplificazione porta ad approssimare i risultati, in quanto anche nodi esterni ai percorsi potrebbero influenzare il risultato, ma con questo ragionamento si dovrebbe arrivare dunque al dover considerare tutto il network esistente, andando a rendere molto lunga e intricata l'analisi.



(a) *Intero network*



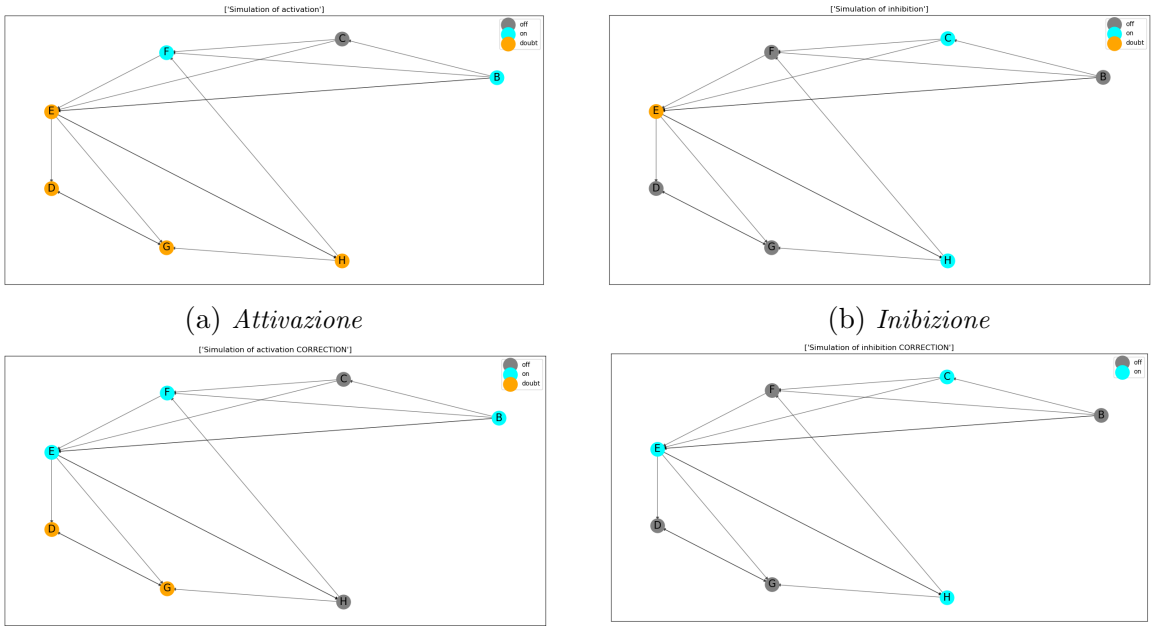
(b) *Taglio del network*

Figura 10.2: *Output grafici rappresentanti il grafo prima e dopo il taglio a seguito della scelta della sorgente 'B' e del target 'G'*

Una volta che si ha il grafo tagliato nella zona d'interesse, vedi Figura 10.2, si procede a considerare due opzioni: il nodo sorgente espresso o inibito. Lo studio parallelo dei due stati iniziali si ripercuote su tutta la rete sottostante; analizzando poi il risultato finale si potrebbe anche avere un'idea del tipo di relazione tra i due nodi d'interesse, se ad esempio i due sono connessi da una rete di attivazione o di inibizione. L'utente è il responsabile della scelta dei due cammini compilando l'apposito campo in seguito a una domanda sulla command window.
A seconda dell'espressione della sorgente, il nome si salva nel vettore dei valori con lo stato uno o zero.
Dopo questa prima inizializzazione dei due vettori, si passa allo studio di tutti i percorsi tra la sorgente e il target, definendo a ogni passo lo start e l'end da cui rispettivamente diparte e arriva l'arco in studio.
Secondo la logica descritta precedentemente dell'espressione del gene e dell'attività dell'arco, si va poi a mettere il risultato all'interno dei due vettori contenenti gli stati pari a uno o a zero.
Dopo aver analizzato tutti i percorsi si tolgono i nodi ripetuti sia nello stato uno che nello stato zero, operazione utile al fine della ricerca dei nodi dubbi che viene fatta con una semplice intersezione tra i due insiemi.
Un nodo dubbio è un nodo che a seconda del percorso si ritrova ad essere sia attivo, stato uno, sia spento, stato zero. Alla fine di tutto ciò si ha una rete dove è stata propagata l'attivazione o l'inibizione, evidenziando su un grafo gli stati dei vari nodi: *on* (ciano), *off* (grigio) e *doubt* (arancione).

Si può osservare in Figura 10.3a la propagazione del valore attivo di espressione della sorgente, mentre in Figura 10.3b è rappresentato il risultato della propagazione dello stato represso della sorgente.

Per ridurre il numero di nodi dubbi è stato fatto un controllo sugli archi entrati su questi nodi.
Infatti tra due nodi possono dipartire più di un arco con diverse azioni, ciò dipende da che ricerca è stata condotta e pubblicata. Questi nodi pertanto non sono sbagliati, dipende da che ricerca si prende in considerazione, per cui possono essere riportati al loro valore aspettato, andando a introdurre altri nodi *ok* nel grafo di confronto. sicuramente questo nodo risulterà corretto, cioè con valore concorde al Golden Standard, a seconda della strada scelta. É possibile osservare la correzione in Figura 10.3c, per l'attivazione, dove i nodi *E* ed *H* passano dall'essere dubbi al divenire *on* e *off* corrispettivamente. In Figura 10.3d si osserva la correzione dell'inibizione del nodo sorgente, ottenendo un risultato privo di nodi dubbi.



(a) *Attivazione*

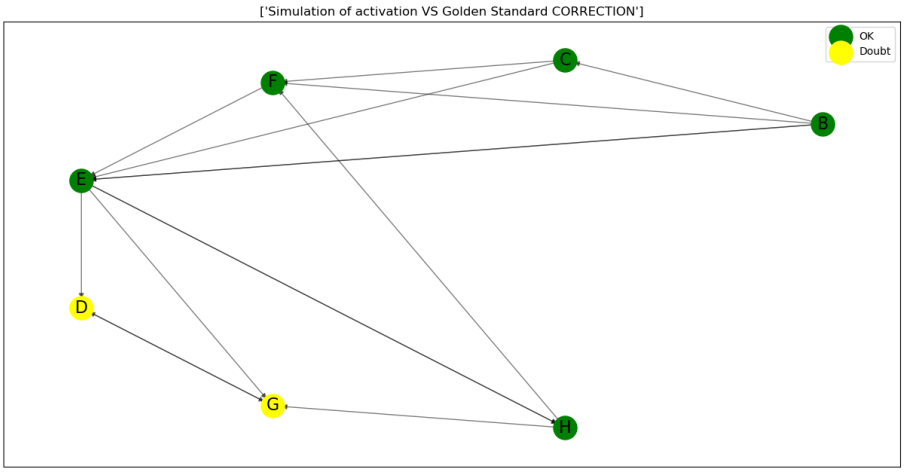
(b) *Inibizione*

(c) *Attivazione corretta*

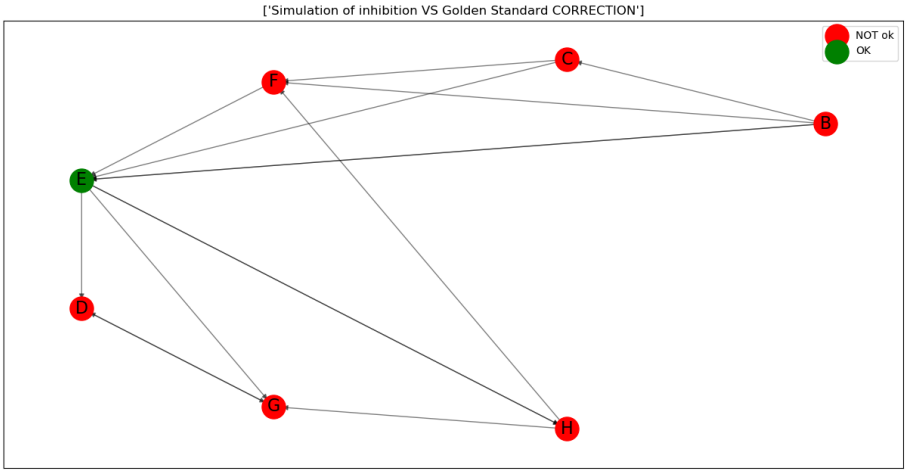
(d) *Inibizione corretta*

Figura 10.3: *Risultati simulazione con propagazione dell'attivazione inibizione del nodo sorgente, prima e dopo la riduzione dei nodi dubbi*

La seconda parte prevede la lettura del file di nodi contenente le varie espressioni, per permettere così un confronto tra quello che è considerato Golden Standard e il grafo risultante dalla propagazione. Leggendo il file nodi si conosce con esattezza il profilo di espressione di ogni nodo, e con un semplice confronto tra liste si arriva alla conclusione di correttezza o inesattezza della propagazione.
Se il valore dei nodi risultano concordanti, quindi etichettati con *ok* (verde), significa che gli archi riportati sono giusti e che l'attivazione o inibizione della sorgente ha portato a risultati positivi. Se invece alcuni elementi vengono catalogati come *not ok* (rosso) significa che c'è stato un problema nella propagazione, e che quindi è possibile che ci sia un problema nella definizione degli archi o dello stato della sorgente. Per i nodi dubbi si lascia il *doubt* (giallo).
Tutto ciò viene riportato all'interno di grafi con legende facilmente leggibili e interpretabili.



(a) *Attivazione*



(b) *Inibizione*

Figura 10.4: *Risultati confronto con Golden Standard dell'attivazione e dell'inibizione della sorgente B*

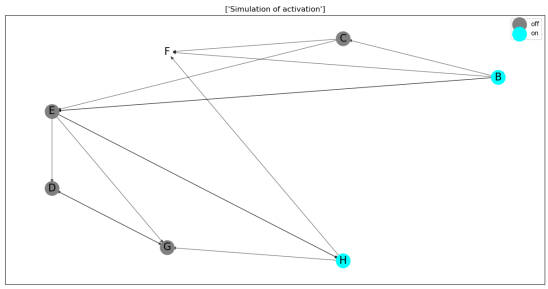
Dalla Figura 10.4 è possibile osservare come la rete sia più compatibile con l'idea di propagazione dell'attivazione del nodo sorgente *B*.

La terza e ultima parte prevede di richiedere all'utente la volontà di cancellare qualche arco, e dunque propagare nuovamente le espressioni con la modifica apportata. Tutto ciò viene riproposto finché l'utente non inserisce più il comando *go*, interrompendo così il ciclo. L'arco che viene eliminato può essere di inibizione o attivazione, a seconda della scelta dell'utente stesso, vedi Figura 10.5.

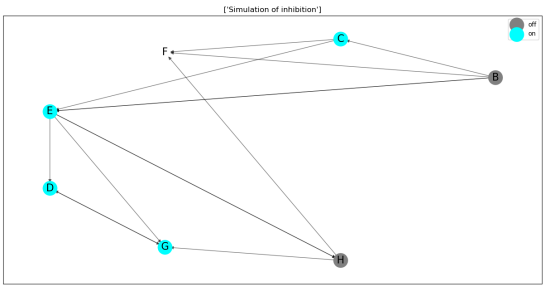
```
Do you want to continue? if yes write 'go', if not write anything: go
What edge do you want to remove? Type source: F
What edge do you want to remove? Type target: E
What edge activity do you want to remove ( -> or -| ): ->
```

Figura 10.5: *Richiesta all'utente sul tipo di arco da eliminare*

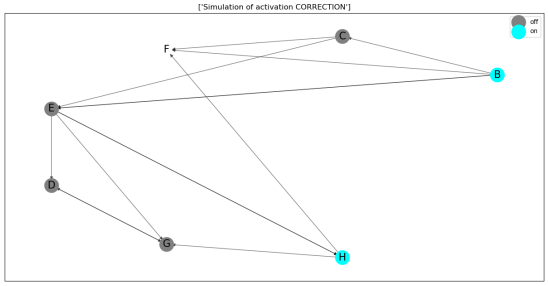
Il codice dopo aver appreso il nodo da eliminare procede con la reiterazione delle funzioni descritte prima. Il tutto viene riproposto e svolto fino a quando l'utente decide di interrompere rispondendo altro che non sia *go* alla richiesta di proseguire.
In Figura 10.6 è possibile vedere l'effetto prodotto dall'eliminazione dell'arco F-E di attivazione (->) sia sulla propagazione dell'espressione della sorgente (figure a sinistra) sia della repressione (figure a destra).



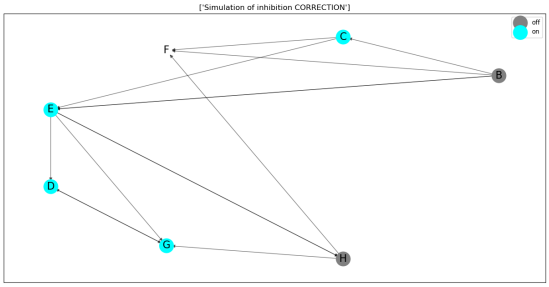
(a) *Propagazione attivazione*



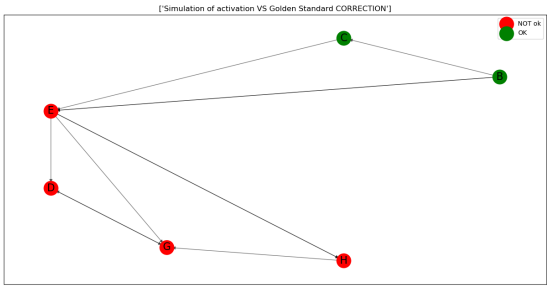
(b) *Propagazione inibizione*



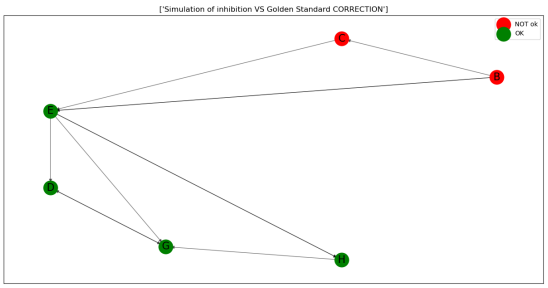
(c) *Propagazione attivazione con meno dubbi*



(d) *Propagazione inibizione con meno dubbi*



(e) *Confronto propagazione attivazione con Golden Standard*



(f) *Confronto propagazione inibizione con Golden Standard*

Figura 10.6: *Risultati reiterazione simulatore senza arco F-E di attivazione (->)*

Capitolo 11

Simulatore d'espressione, con produzione di un Report di coerenza degli archi

Prototipo di tool di Validazione della Rete, mediante il forzamento dell'espressione di un pool di geni

Il simulatore progettato ed illustrato in questo capitolo si prefigge lo scopo di restituire, forzando lo stato d'espressione di un determinato pool di geni, il grado di coerenza della rete in relazione ad esso, sottoforma di percentuali di correttezza, errore ed incertezza relative ad ogni arco elaborato.
I file utilizzati per lo sviluppo sono sempre due e sono analoghi a quelli riportati nel Capitolo 10, in Figura 10.1. L'unica e sostanziale differenza, che rende questo simulatore più flessibile rispetto al prototipo visto nel Capitolo 10, è che il file contenente l'espressione dei nodi non è limitato a dover rappresentare forzatamente l'espressione totale di tutti i nodi presenti nella rete, bensì può contenere semplicemente l'espressione di un sottoinsieme di essi.

Risulta infatti difficile ipotizzare di poter avere il profilo d'espressione completo della rete in analisi, per via della continua e necessaria evoluzione degli studi biologici, che rendono ogni informazione attuale incerta in vista delle future, e delle differenze sostanziali dipendenti dai protocolli di conduzione dei vari esperimenti. Pensare di avere, invece, l'espressione di un ristretto pool genico e tentare di ottenere da essa valutazioni sulla rete totale, risulta essere una possibilità molto più allettante e scalabile.
Successivamente alla lettura dei due file e, quindi, alla costruzione della rete e alla definizione dello stato dei geni posseduto, per ognuno dei geni con espressione nota si effettuano due operazioni:

- Calcolo di tutti i Simple Paths aventi esso come sorgente e, iterativamente, tutti gli altri nodi del grafo come destinazione, salvati in una lista chiamata *paths_from*;
- Calcolo di tutti i Simple Paths aventi, iterativamente, ogni nodo del grafo come sorgente ed esso come destinazione, salvati in una lista chiamta *paths_to*.

La necessità di calcolare tutti i percorsi entranti ed uscenti nei geni d'interesse risiede nella volontà di avere un'indicazione di tutte le cause e di tutti gli effetti che possono derivare da quell'espressione. Non risulta difficile comprendere che l'espressione di un singolo gene può essere sia frutto di diversi percorsi regolatori, che essere la causa, a sua volta, di una cascata di effetti. Ciò può essere ottenuto, dunque, elaborando tutti i percorsi entranti ed uscenti da questi geni.

In seguito alla semplice identificazione dei percorsi e al salvataggio nelle corrispettive liste, queste ultime vengono rese uniche. Il grafo su cui si opera, infatti, risulta essere un multi-grafo direzionato e, come detto nel Capitolo 4, questa tipologia di grafi permette di avere due o più archi tra due nodi, differenziandoli usando attributi. Il calcolo dei percorsi, però, non permette di visualizzare gli attributi di ogni arco, restituendo semplicemente, per ogni arco multiplo, lo stesso percorso ripetuto tante volte quante sono gli archi tra due nodi. I percorsi vengono così resi unici e salvati in due nuove liste: *paths_to_unique, paths_from_unique*.

Successivamente, la lista *Res*, che conterrà la valutazione di correttezza di ogni arco, viene inizializzata. Contemporaneamente a questa operazione, anche la lista *count* contenente l'attributo *int_count* per ogni arco valutato, che come già detto indica quanti archi di quella tipologia esistono nel grafo.
Ogni arco, contenuto in *Res*, può avere 3 tipi di Valutazione:

- 0, che indica la correttezza dell'arco;
- 1, che ne indica l'errore;
- 2, che indica un'incertezza nella considerazione dell'arco; saranno illustrati successivamente i motivi che hanno portato all'esistenza di questa valutazione intermedia.

Si effettuano, a questo punto, due elaborazioni principali:

1. Ogni percorso uscente dai geni d'interesse viene analizzato e viene su di esso propagata l'espressione del gene di partenza, che risulta nota e salvata nella variabile *STAT_PREC*, contenente l'espressione di ogni gene di partenza dell'arco, mentre la variabile *PREC*, rappresentante la valutazione dell'arco precedentemente analizzato viene inizializzata a 0, che indica, come già detto, la correttezza.
   La propagazione viene così attivata e per ogni arco analizzato, tenendo conto di eventuali archi multipli, che implicano propagazioni di effetti diverse dipendentemente dall'attributo direction dei diversi archi paralleli, nonché della valutazione dell'arco precedente, si inizializzano e si aggiornano successivamente 3 variabili liste principali:
   - *STAT*, che indica lo stato del gene di destinazione dell'arco, ottenuto analizzando sia l'effetto di regolazione del singolo arco considerato, indicato proprio dall'attributo direction, sia lo stato del gene di partenza;
   - *DIRECTION*, contenente l'azione effettuata da quell'arco, definita nell'omonimo attributo di esso;
   - *EVAL_PREC*, indicante la valutazione dell'arco precedente a quello appena analizzato ed utilizzata per indicizzare in maniera univoca gli stati elaborati alle valutazioni del/degli arco/archi precedente/i ad esso.

Queste liste vengono successivamente analizzate, al fine di restituire una valutazione di correttezza dello stesso. Si verifica, prima di tutto, se il nodo di destinazione dell'arco appena analizzato ha espressione nota, dunque definita in fase iniziale, o meno. Se il nodo ha espressione nota, si confronta l'espressione ottenuta con la regolazione al valore noto e si effettua la valutazione:

- I due profili d'espressione non combaciano, dunque se l'arco analizzato precedentemente risulta corretto, l'arco viene valutato come sbagliato; se l'arco analizzato precedentemente, invece, risulta sbagliato o dubbio, l'arco viene valutato come dubbio. Questa scelta sviluppativa è dovuta al fatto che, essendosi già verificato un errore all'interno del percorso, non è possibile stabilire se quell'arco risulta essere sbagliato o corretto, in quanto strettamente dipendente dall'espressione di un arco precedente che risulta incongruente;
- I due profili d'espressione combaciano, dunque se l'arco analizzato precedentemente risulta corretto, l'arco viene valutato come corretto; se l'arco analizzato precedentemente risulta sbagliato o dubbio, anche quest'arco, nonostante restituisca un profilo coerente, viene segnalato come dubbio, poiché è strettamente dipendente da un errore verificatosi precedentemente.

Se il nodo, invece, non ha espressione nota, non potendo fare ulteriori confronti, l'arco viene valutato come corretto, a meno che la precedenza non sia dubbia o errata; in tal caso l'arco viene valutato come dubbio.
In questo modo è possibile ottenere una valutazione dell'arco singolo, ma strettamente correlata alla storia del percorso di estrazione dello stesso.
Ogni valutazione viene inserita sotto forma di lista contenente l'arco, l'azione mediata dallo stesso e la valutazione, all'interno della lista *Res*; contemporaneamente a questa azione, si aggiornano i valori degli Stati precedenti e delle precedenze con quelli appena valutati e si prosegue l'analisi, che risulterà quindi strettamente correlata sia alla storia precedente del percorso, sia al numero di archi multipli intercorrenti tra le due entità analizzate. Questa operazione permette di indagare gli effetti di ogni percorso ed arco regolatorio.

2. La stessa operazione viene effettuata sui percorsi entranti nei geni d'interesse, svolgendo però ogni operazione a ritroso; si inizia, dunque, l'elaborazione di ogni percorso a partire dall'ultimo gene, che risulta avere espressione nota, analizzando le cause di ogni percorso ed arco regolatorio.

A questo punto, a partire dalla lista contenente tutte le valutazioni per ogni arco analizzato relativamente al percorso di estrazione dello stesso, si prosegue con la generazione di una lista *Report*, contenente per ogni arco di uno specifico tipo (quindi per ogni tripla di tipo sorgente, destinazione, azione mediata), 3 valori percentuali:

- **Correct Ratio (%)**, rappresentante il rapporto tra il conteggio delle valutazioni corrette di quell'arco e il conteggio delle valutazioni totali dello stesso, espresso in valore percentuale;
- **Wrong Ratio (%)**, rappresentante il rapporto tra il conteggio delle valutazioni errate di quell'arco e il conteggio delle valutazioni totali dello stesso, espresso in valore percentuale;
- **Doubt Ratio (%)**, rappresentante il rapporto tra il conteggio delle valutazioni dubbie di quell'arco e il conteggio delle valutazioni totali dello stesso, espresso in valore percentuale;

Successivamente è stato assegnato anche uno stato Predetto ad ogni arco, sfruttando delle soglie definite in fase di sviluppo, in base ad osservazioni fatte sulle reti analizzate, basate sui vari valori percentuali ricavati. L'assegnazione dello stato viene svolta in questo modo:

- *Correct*: $\begin{cases} CorrectRatio \geq 2WrongRatio \\ CorrectRatio + DoubtRatio \geq 4WrongRatio \end{cases}$
- *Wrong*: $\begin{cases} WrongRatio \geq 2CorrectRatio \\ WrongRatio + DoubtRatio \geq 4CorrectRatio \end{cases}$
- *Doubt*: in tutti gli altri casi

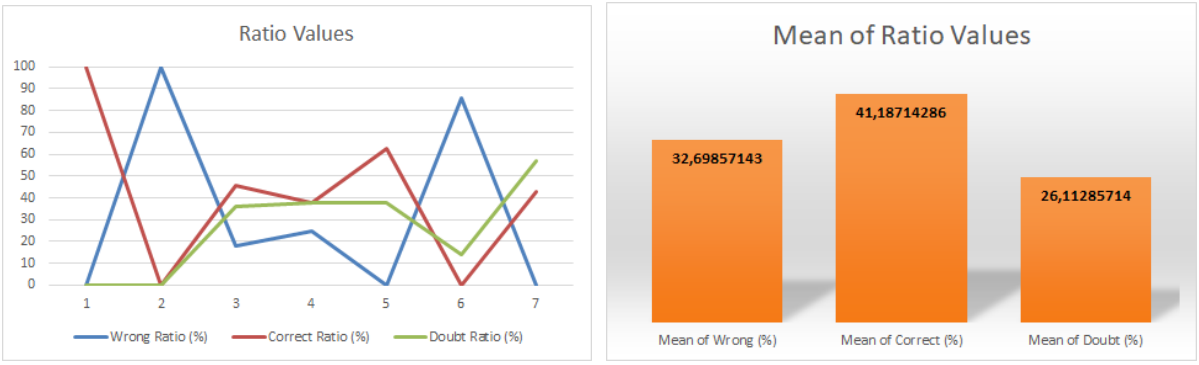
Ovviamente lo stato predetto e la sensibilità predittiva sono strettamente correlate alle soglie definite, che possono quindi essere modificate e calibrate in base alle proprie necessità.

Il Report così definito è stato dunque salvato in un file Excel, analogo a quello esemplificativo mostrato in Figura 11.1.

	1: Edge	2: Action	3: Wrong Ratio (%)	4: Correct Ratio (%)	5: Doubt Ratio (%)	6: Predicted Status
0	('B', 'C')	->	0	100	0	Right
1	('B', 'C')	-|	100	0	0	Wrong
2	('B', 'D')	->	18,18	45,45	36,36	Right
3	('D', 'E')	->	25	37,5	37,5	Doubt
4	('D', 'E')	-|	0	62,5	37,5	Right
5	('E', 'F')	->	85,71	0	14,29	Wrong
6	('A', 'B')	->	0	42,86	57,14	Right

Figura 11.1: *Report Generato dal prototipo di tool di Validazione*

È stata implementata anche una doppia rappresentazione grafica dello stesso, mediante l'utilizzo di *Microsoft Excel*, tramite un line-chart rappresentante sull'asse delle ordinate l'andamento dei 3 Valori Percentuali illustrati precedentemente, per i diversi archi, riportati sotto forma di indici numerici sull'asse delle ascisse, e un istogramma, rappresentante le medie dell'andamento dei 3 valori percentuali, utilizzate per valutare le performance totali della rete. I due grafici sono mostrati rispettivamente in Figura 11.2a e 11.2b.

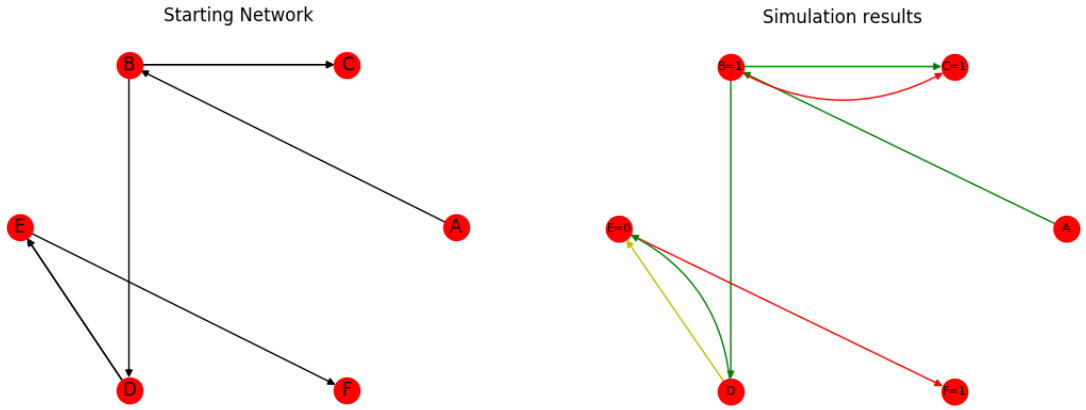


(a) *Andamento dei Valori di Correttezza/Errore/Incertezza percentuali*

(b) *Istogramma delle Medie dei Valori di Correttezza/Errore/Incertezza percentuali*

Figura 11.2: *Rappresentazioni Grafiche del Report*

Anche una versione grafica del processing sul grafo è stata sviluppata, la quale permette di visualizzare ogni arco colorato secondo lo stato predetto. I colori permettono dunque di rappresentare i vari stati:
Corretto (verde), Sbagliato (rosso), Dubbio (giallo). In Figura 11.3a e 11.3b sono mostrate, rispettivamente, la rete iniziale e la rete post-processata. Va segnalato che non è stato possibile apporre una label ad ogni arco, indicante l'attributo direction e, quindi, l'azione dello stesso, per il quale si rimanda al file contenente il Report.



(a) *Network iniziale*

(b) *Rappresentazione del network elaborato tramite il prototipo di tool di Validazione*

Figura 11.3: *Network esemplificativi*

Versioni Definitive

Come detto nei capitoli precedenti, TheRingDB risulta essere una rete eterogenea e multilivello, che non limita il campo d'azione ad uno solo dei network visti nel capitolo 4, ma li unisce, consentendo un'analisi combinata dei vari aspetti del problema trattato.
Per tale motivo sono state implementate 4 diverse versioni definitive del tool:

1. **Simulatore d'espressione della GRN**
   Questo simulatore effettua un processing della iniziale della rete analizzata, finalizzato all'eliminazione degli archi rappresentanti interazioni Proteiche (PPI) e ogni altra interazione non finalizzata alla regolazione e alla trasmissione dei segnali nella rete genica. In questo modo, è possibile limitare il campo d'interesse della rete e il focus del simulatore, passando da un'analisi di tipo multilivello, riguardante diverse tipologie e sorgenti di dati, ad un'analisi incentrata sulla sola Gene Regulatory Network.

2. **Simulatore d'espressione con PPI monostato**
   Questa tipologia di simulatore non effettua alcun processing sulla rete, tenendo conto di tutte le diverse interazioni presenti dalle entità. Ciò ha richiesto un'opportuna considerazione degli archi rappresentanti le interazioni tra le proteine e una strategia di collegamento tra i due problemi che dovrebbero essere condotti in maniera parallela. Esso, infatti, propaga semplicemente lo stato tra le due entità che formano il complesso proteico, partendo da un gene con espressione precedentemente processata, permettendo così di considerare solo due condizioni:

   - *Attivazione Totale*: Il complesso assume uno stato di tipo 1-1, propagando l'accensione dal primo nodo al secondo. Ciò indica che le proteine cooperano e che l'interazione è attiva, poiché sono entrambe espresse, indicando, in maniera indiretta, l'attivazione di entrambi i geni che le producono.
   - *Inibizione Totale*: Il complesso assume uno stato di tipo 0-0, propagando l'accensione dal primo nodo al secondo. Ciò indica che le proteine non cooperano e che l'interazione non è attiva, poiché sono entrambe non espresse, indicando, in maniera indiretta, la disattivazione di entrambi i geni che le producono.
     Questo è un assunto che è stato fatto in fase di sviluppo, limitando il campo di interesse del simulatore, ma va segnalato che questa considerazione non tiene conto di svariate possibilità che potrebbero provocare una mancata interazione di proteine espresse da geni in realtà attivi: le proteine, infatti, potrebbero in realtà essere codificate in maniera scorretta, a causa di modifiche post-trascrizionali, mutazioni e alterazioni nella conformazione strutturale, nel sito di fosforilazione o nel binding site, e quindi non interagire; altre possibilità non contemplate sono quelle dell'eventuale "silenziamento" delle proteine da parte di miRNA, di processi degradativi delle stesse, o di una mancata localizzazione.

Una considerazione di questo tipo consente, a partire dalle interazioni proteiche, di fare valutazioni indirette sullo stato di espressione dei geni dalle quali sono prodotte. Il gene analizzato, con espressione così valutata, potrebbe avere però espressione nota, forzata in fase iniziale e i due valori, come già visto, potrebbero contrastare o essere corrispondenti. Nessuna valutazione di correttezza viene ad ogni modo effettuata, consentendo semplicemente la propagazione delle catene di cause ed effetti.

3. **Simulatore d'espressione con PPI monostato valutata**
   Questa tipologia di simulatore è analoga alla precedente, implementando le stesse strategie di sviluppo e considerando le stesse condizioni.
   Questa versione, però, effettua la valutazione della correttezza tra lo stato d'espressione del gene, ricavato indirettamente a partire dalle interazioni proteiche, e un eventuale stato d'espressione noto dello stesso, pur non inserendola nel report, modificando così i valori degli stati precedenti degli archi in fase di processing.
   In questo modo, ogni eventuale errore presente nelle valutazioni indirette non viene segnalato come tale, in relazione al Report finale, poiché strettamente dipendente dalle considerazioni e dalle assunzioni fatte in fase di sviluppo, pur mostrandone la presenza in un eventuale incremento del numero degli archi regolatori dubbi, in virtù del fatto che un eventuale errore risultante dalla valutazione di correttezza di tale tipo di interazioni, si potrebbe propagare successivamente nei cammini regolatori, generando un numero di dubbi maggiore rispetto al caso precedente.
   Così facendo, è come se le due versioni di simulatore si differenziassero nel marcare o meno l'effetto dovuto alla presenza di interazioni proteiche.

4. **Simulatore d'espressione con PPI bistato valutata**
   Questa tipologia di simulatore non effettua alcun processing sulla rete, tenendo conto di tutte le diverse interazioni presenti dalle entità, richiedendo nuovamente strategie di collegamento tra le diverse tipologie di interazioni. Esso, differentemente dal caso precedente, in fase di analisi delle interazioni proteiche, partendo dalla prima entità, con stato noto, propaga ambedue gli stati possibili (attivo e disattivo), simulando così una non predicibilità del valore. In questo modo sono 3 le condizioni considerate:

   - *Attivazione Totale del complesso*: Il complesso assume uno stato di tipo 1-1, propagando l'accensione dal primo nodo al secondo. Ciò indica che le proteine cooperano e che l'interazione è attiva, poiché sono entrambe espresse, indicando, in maniera indiretta, l'attivazione di entrambi i geni che le producono.
   - *Inibizione Totale*: Il complesso assume uno stato di tipo 0-0, propagando l'accensione dal primo nodo al secondo. Ciò indica che le proteine non cooperano e che l'interazione non è attiva, poiché sono entrambe non espresse, indicando, in maniera indiretta, la disattivazione di entrambi i geni che le producono.
   - *Attivazione/Inibizione parziale*: Il complesso assume uno stato di tipo 1-0/0-1 (Dipendentemente dallo stato dell'entità iniziale). Ciò indica che le proteine non cooperano e che l'interazione non è attiva, poiché solo una di loro è espressa, indicando, in maniera indiretta, l'attivazione del corrispondente gene e la disattivazione dell'altro.

Implementando questa strategia è come se non si facessero valutazioni a priori; inoltre, sebbene la non esistenza della proteina risulti essere sempre algoritmicamente associata alla mancata espressione del gene, ignorando tutte le altre possibilità elencate precedentemente, l'assegnazione di entrambi gli stati alla proteina e, in maniera indiretta, al gene codificante, permette, ad ogni modo, di vagliare le possibilità associate ad entrambi gli stati di espressione del gene coinvolto nel complesso. Ovviamente, al guadagno di capacità di generalizzazione del sistema, dato da questo assunto, corrisponde una mancata specificità da parte dello stesso.

Anche in questo caso è stata effettuata la valutazione di correttezza, sebbene non riportata nel Report, come visto nel caso precedente.

Ovviamente, ammettendo per ogni tipo di complesso proteico 2 possibili stati, qualora si generassero degli errori in fase di valutazione, il numero di dubbi crescerebbe ancor di più, poiché è come se ogni percorso venisse valutato almeno due volte (anche più, in caso di più archi di interazione proteica appartenenti ad esso), marcando gli effetti delle interazioni proteiche ma pur rendendoli difficilmente distinguibili e, quindi, aumentando la possibilità di rendere i dati ottenuti di difficile interpretazione.

Parte IV

Applicazione del simulatore a uno specifico pool di geni

Il simulatore definitivo è stato applicato nello studio di determinati geni analizzati dai ricercatori di scienze mediche - genetica medica dell'Università di Torino.
In particolare è stata fornita una lista contenente 126 nomi di entità biologiche che interagiscono differenzialmente con il gene *TLK2* quando questo è mutato, vedi Tabella 11.1.

Le interazioni individuate sono sia dirette che indirette, in quanto è stato utilizzato il metodo *BioID* appartenente alla tecnica *PDL - proximity dependent labeling* [41] [38].
Gli studi attraverso l'utilizzo della PDL si basano sull'identificazione delle PPI, protein protein interaction, in vivo; e vengono catturate sia le proteine che interagiscono direttamente sia quelle che si trovano in prossimità della zona d'interesse.
Il PDL sfrutta enzimi capaci di modificare proteine grazie alla catalisi tra gruppi reattivi dell'enzima stesso e i target amminoacidici delle proteine.
In particolare il BioID sfrutta il gruppo reattivo biotin-ligasi del BirA, enzima estratto dall'*E.coli*, nel caso della prima versione; mentre la versione successiva, la BioID2, prevede l'utilizzo della biotin-ligasi isolata dall'*aquifex aelicus*.
La BioID ha dimensioni ridotte rispetto la BioID, circa un terzo, perciò prevede l'utilizzo di meno biotina per ottenere gli stessi risultati della prima. Per riuscire a catturate proteine di grandezza maggiore rispetto all'enzima o addirittura complessi molecolari, si usano dei linker flessibili per aumentare il raggio di labeling che corrisponde a 10-15mm.
La biotin-ligasi si lega a una sequenza di consenso della proteina, ma solo se l'enzima presenta un centro reattivo. Il centro reattivo infatti permette di poter legare i residui amminici della lisina contenuta nella proteina all'enzima BirA, dando origine a legami amide e al trasferimento della biotina dalla specie reattiva alla proteina.
Il centro reattivo del BirA si origina ingegnerizzando la molecola e dunque mutando un solo amminoacido, l'R118G, ottenendo così la specie reattiva BirA*.
La specie BirA* è molto instabile e tende a modificare tutte le proteine nel suo raggio d'azione, sia quelle con cui ha interazione diretta sia quelle in prossimità.
Le proteine modificate dal BirA, e dunque che hanno subito il trasferimento della biotina, vengono successivamente catturate da sfere di agarosio rivestite di streptavidina e poi analizzate con spettrometria di massa [41] [38].

Per problemi computazionali sono stati introdotti dei parametri di semplificazione per la produzione della sotto-rete di studio. É stato applicato un taglio di un fattore due nella generazione dei *simple paths* tra *TLK2* e gli altri geni, ottenendo in questo modo collegamenti tra le due entità aventi massimo 3 nodi. Procedendo in questo modo si sono ottenuti diversi file *pickle*:

- uno contenente il grafo completo di *TLK2* e gli elementi della Tabella 11.1.
- altri aventi i cammini *TLK2*-elemento singolo della Tabella 11.1.

DIDO1	PPP1R10	SBNO1	GTF2A1	CHD8	PPP4R3A
SUGP1	SCML2	TPR	RAD50	ZNF318	PRPF3
SAP30BP	SPDL1	ARID3B	CTNNBL1	RFC1	POLD1
NCBP1	PRIM1	SUPT5H	ZNF148	NZNF281	DHX8
MLH1	KDM3A	WRNIP1	ZCCHC8	TASOR	BRD4
JMJD1C	XAB2	PAXBP1	BRIP1	RNF40	ORC2
PES1	BOP1	MSH3	ZBTB10	LYPLA2	CDK7
LIG1	MEF2D	ARID3A	METTL3	MTHFD2	PMS1
SFSWAP	SMARCA1	TLK1	ZHX3	NACC1	NELFB
ZNF131	GPATCH1	CACTIN	ASF1B	ASF1A	CHD7
DHX35	NELFCD	MYBL2	ZNF362	CIC	BEND3
MSANTD2	BAZ2A	UBN1	DYNLL1	ISY1	SMU1
ARNT	AQR	RSBN1L	MLLT1	ZSCAN18	PMS2
LIN9	PAPOLG	ELF2	ZKSCAN4	ZNF512B	BUD31
E2F3	ELF1	POU2F1	ZEB1	ZBTB9	CCDC174
NR2C1	SIX4	NAB1	HIST1H2AA	FLI1	ALX4
TFAP4	NFIX	NFIC	GATD3B	NFIA	JUNB
EBF3	JUND	LIN52	TPGS1	SIN3B	MBD2
KLHL13	SALL3	RPS6KA5	POLH	HIRA	HDAC5
RACGAP1	TADA3	TAF5L	USP28	RNF8	YEATS2
KNL1	NFAT5	PHF21A	RBM20	PHF12	MBIP

Tabella 11.1: *Elenco dei nomi dei 126 elementi oggetto di studio al dipartimento di medical sciences - medical genetics dell'Università di Torino*

La riduzione drastica del grafo in questo modo ha portato alla produzione di 123 *pickles* appartenenti alle coppie di geni, in quanto il gene *GATD3B*, il transcription factor *TASORS* (conosciuto anche con l'alias *FAM208A*) e l'entità *NZNF281* risultano al di fuori della soglia scelta.
Tutte le coppie formate da *TLK2* e dalle 123 entità presenti sono state studiate andando ad analizzare i risultati del simulatore imponendo il nostro elemento cardine *TLK2* sempre espresso e l'altro componente della coppia prima espresso e poi represso.

Successivamente è anche stata analizzata con più accuratezza una rete avente quattro geni d'interesse: *TLK2*, *TLK1*, *ASF1A*, *ASF1B* più alcuni nodi presenti nei cammini di collegamento tra questi. Il motivo della scelta dei geni d'interesse è basato su una ricerca di relazioni consolidate in letteratura con l'elemento d'interesse *TLK2*.

Il gene *TLK2* a capo della lista ottenuta, che si traduce nell'uomo in enzima *serine/threonine-protein kinase tousled-like 2* [35], interagisce in modo consistente con altri tre geni presenti nella Tabella 11.1 [18] [17]:

- *TLK1* [34];
- *ASF1B* [28];
- *ASF1A* [27].

La particolare interazione tra questi quattro oggetti è stata approfondita con la costruzione di una sotto-rete, partendo dalla rete del *pickle* avente i *simple paths* con taglio di un fattore due, composta dalle quattro entità e dai nodi presenti nei cammini di congiunzione tra esse.

Il motivo per il quale si sono scelti i tre elementi tra tutti quelli in elenco, è la loro stretta correlazione con il gene cardine *TLK2*.
L'importanza dello studio sul gene *TLK2* è la sua consolidata correlazione con disturbi del neurosviluppo, come confermato da numerose ricerche e dalla sua presenza nell'*Autism database*, un database che racchiude i geni connessi all'autismo.
Il gene *TLK2* è espresso in tutti i tessuti, compreso il tessuto cerebrale del feto [40]. Presenta un picco di attività durante la fase di sintesi del ciclo cellulare e nel corso della separazione dei due filamenti del DNA viene brevemente inibita la sua attività; tutto ciò lo rende strettamente collegato alla regolazione della fase di replicazione del DNA e dunque molto delicato. Inoltre è stato scoperto che i Transcription Factors *ASF1* sono i suoi substrati fisiologici, rendendo il *TLK2* partecipe anche dell'assemblaggio della cromatina [40].

In uno studio condotto da Reijnders et al. denominato *"De Novo and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype-Phenotype Evaluation of a Distinct Neurodevelopmental Disorder"* [40], sono state analizzate più varianti del gene TLK2 e la loro incidenza nella modifica dell'aspetto facciale in soggetti affetti dal disturbo dello spettro autistico.
Le varianti del gene sono state scansionate in diversi database, dando un esito positivo per alcune e le altre invece sono state ricercate all'interno del genoma dei genitori. Questa scansione ha evidenziato la presenza di varianti del gene con una funzione ridotta nelle madri, le quali oltre a presentare un certo ritardo nel linguaggio e nello sviluppo e un disturbo bipolare, hanno mostrato una certa somiglianza con i figli per quanto riguarda il fenotipo facciale.
Nel presente studio sono stati anche condotti degli studi su caratteristiche associate alle varianti del *TLK2*, come si può vedere in Figura 11.4.
Una completa perdita delle funzioni di *TLK2* è stata anche associata all'arresto della divisione nucleare e conseguente apoptosi delle cellule.

Il *TLK1* è collegato al *TLK2* grazie alla loro similarità; infatti anche se i due sono presenti in diversi cromosomi, hanno l'84% dell'identità a livello proteico, rendendo i due cammini di espressione molto simili, e il 96% a livello del dominio di chinasi, rendendo le funzioni praticamente identiche.
Questa similitudine fa sì che molto spesso vengono identificati entrambi con il termine *TLKs*.
La presenza all'interno degli organi è pressoché ben distribuita per quanto riguarda il *TLK1*, mentre per il *TLK2* è stata ritrovata una grande concentrazione nei testicoli, prendendo parte alla meiosi maschile e rendendo possibile la spiegazione della maggior incidenza dell'ASD sui maschi rispetto alle femmine; comunque anch'esso si ritrova ben distribuito negli altri organi [44].

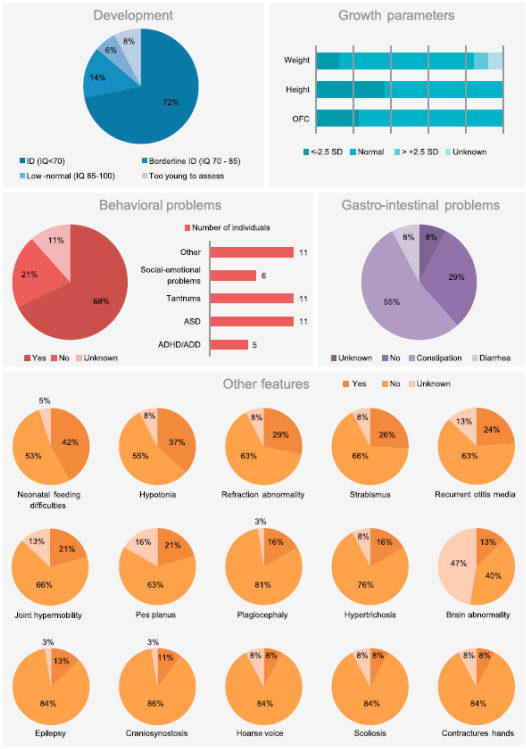


Figura 11.4: *Grafici rappresentanti lo spettro clinico associato alle mutazioni del gene TLK2. Fonte "De Novo and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype - Phenotype Evaluation of a Distinct Neurodevelopmental Disorder" di Reijnders et al.[40]*

Secondo lo studio *"The Tousled-like kinases regulate genome and epigenome stability: implications in development and disease"* di Segura-Bayona e Stracjer [43], i due *TLK* hanno un ruolo importante nella replicazione, riparazione e trascrizione del DNA, e anche nello stesso sviluppo dell'organismo, in quanto presentano un picco di funzionalità nella fase S del ciclo cellulare.
La similarità tra *TLK1* e *TLK2* sta anche nella loro influenza nello sviluppo embrionale. É stato già visto come una perdita di funzionalità del *TLK2* si manifesti con disturbi del neurosviluppo e della sindrome autistica, e la sua totale deplezione porti a un fallimento placentale con conseguente morte dell'embrione; per quanto riguarda invece il *TLK1* è stato studiato che una deplezione porta a una downregolazione di geni pluripotenti e di differenziazione di cellule staminali nell'embrione, oltre a riattivare virus latenti come il virus Epstein-Barr.
Lo studio si conclude affermando che non è escluso che i *TLK* regolino lo splicing microesonale, funzione importante poiché una sua misregolazione è strettamente correlata al disturbo dello spettro autistico.

Secondo lo studio *"Tousled-like kinases phosphorylate Asf1 to promote histone supply during DNA replication"* di Klimovskaia et al. [22], durante la fase di sintesi c'è una richiesta di istoni per ripristinare l'organizzazione della cromatina del DNA appena sintetizzato, e l'elemento chiave alla base dell'induzione di produzione di istoni è la fosforilazione dell'elemento *ASF1* da parte dei *TLKs*.
Durante la fase di sintesi della cellula, il DNA viene duplicato e la sua organizzazione in cromatina deve essere riprodotta affinché vengano mantenute le informazioni genetiche ed epigenetiche in ognuna delle due cellule figlie.

Per mantenere la densità del nuclosoma agli istoni della cellula madre devono essere aggiunti nuovi istoni sintetizzati, gestiti da proteine chaperon. Ogni nucleosoma è composto da due ripetizioni dei seguenti istoni:H2A, H2B, H3 e H4 [33]. L'eterodimero H3-H4 ha come chaperon la proteina *ASF1*, che nell'uomo è presente nelle isoforme *ASF1a* e *ASF1b*. Nello specifico, l'isoforma *ASF1b* è presente solo durante la proliferazione cellulare, mentre l'isoforma *ASF1a* è presente anche durante la quiescenza delle cellule, quindi quando queste sono a riposo.
Lo studio di Klimovskaia et al. [22] esegue numerose osservazioni sulla correlazione tra i *TLKs* e gli *ASF1*, identificando ben quattro siti in cui avviene la fosforilazione dell'*ASF1* da parte dei *TLKs*. Questa fosforilazione, e dunque l'attivazione della richiesta di istoni, è molto importante perché facilita la progressione della fase di sintesi cellulare, tanto che se le isoforme *ASF1a* e *ASF1b* sono assenti solo il 50% delle cellule entra nella fase-S.
Lo studio si conclude analizzando un basso livello di funzioni dei *TLKs* che portano a difetti della cromatina, evidenziando l'importanza della funzione di fosforilazione durante la replicazione e la richiesta di istoni, vedi Figura 11.5. Inoltre una totale assenza di *ASF1* si ha una morte del topo knockout.
Mutazioni del *TLK2* hanno impatto nella regolazione del *ASF1a* e dell'istone *H3*, che si traducono in problemi dello splicing microesomico neuronale, strettamente correlato con il disturbo della sindrome autistica [43].

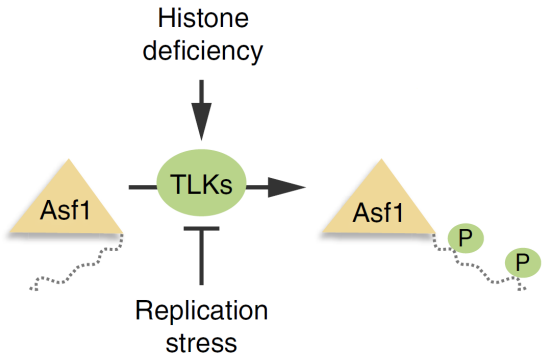


Figura 11.5: *Modello di fosforilazione dell'ASF1 da parte del TLK a seguito di una richiesta di istoni. Fonte "Tousled-like kinases phosphorylate Asf1 to promote histone supply during DNA replication" di Klimovskaia et al.[22]*

Bisogna specificare tuttavia che i collaboratori specializzati in scienze mediche - genetica medica dell'Università di Torino, hanno svolto studi non sull'espressione degli elementi contenuti nella Tabella 11.1 e del gene *TLK2*, bensì sull'aumento o riduzione dell'interazione rispetto al wildtype sulle due variazioni missense. Le variazioni missenso sono mutazioni che hanno come effetto la codifica di un amminoacido diverso da quello originario, e di conseguenza producono effetti sull'interazione con gli altri geni.
Essendo dati riferiti all'interattoma e non all'espressione vera e propria, è stato deciso di assumere che l'aumento dell'interazione degli elementi corrisponda alla loro stessa sovraespressione, viceversa una riduzione alla sotto-espressione. Tutto ciò è stato effettuato per il mero fine di poter costruire il simulatore basato sulle analisi delle espressioni in mancanza dei dati necessari.

Capitolo 12

Setup e osservazioni

Ogni tipologia di simulatore, riportata nel Capitolo 11, è stata utilizzata, al fine di produrre i risultati finali.
La rete è stata opportunamente filtrata e processata, come sarà illustrato nei paragrafi successivi, al fine di allinearla agli obiettivi da raggiungere.
In base alle necessità del problema e ai risultati ottenuti, sono state implementate 5 micro-strategie di Simulazione:

- Simple Paths Simulation;
- Subnetwork Simulation;
- Single-Edge Deletion Simulation;
- Clustered-Edges Total Deletion Simulation;
- Clustered-Edges 2-Combinations Deletion Simulation.

In fase progettuale è stato stabilito se e come applicare, per ogni simulatore, le diverse simulazioni, ottenendo così risultati eterogenei e differenziati.

Simulatore d'espressione della GRN

Come previsto dal simulatore utilizzato, ogni rete è stata inizialmente filtrata e processata, al fine di eliminare tutte le interazioni proteiche presenti, riconducendo il problema analizzato ad una analisi della sola rete regolatoria.
Per questa configurazione, sono state effettuate solo due tipologie di simulazioni:

- Simple Paths Simulation;
- Subnetwork Simulation;

• **Simple Paths Simulation**
Questa tipologia di simulazione è stata condotta utilizzando come sorgente, dalla quale estrarre la rete, il file .gpickle contenente tutti i simple-paths da e verso *TLK2*, prendendo come seconda sorgente/destinazione ognuno dei singoli geni di nostro interesse. Non è stato necessario modificare ulteriormente la rete per ridurre i costi computazionali.
Ogni simulazione è stata condotta considerando *TLK2* sempre espresso, mentre gli altri geni sono stati forzati sia in condizione di espressione che inibizione; dunque ogni rete è stata simulata 2 volte, per un totale di 6 simulazioni.
La motivazione alla base di questo tipo di simulazione è dovuta alla mancanza di dati sull'effettiva espressione dei geni di nostro interesse. Infatti, pur supponendo *TLK2* come espresso, in quanto una sua inibizione non permetterebbe di saggiare gli effetti apportati da una sua mutazione, non è possibile definire a priori lo stato di espressione degli altri geni, in condizioni ottimali, se non facendo riferimento agli studi di letteratura effettuati.
Questa tipologia di simulazione si prefigge dunque lo scopo di identificare il *"best status"* di espressione dei geni analizzati, in condizioni di normalità, in virtù di una valutazione delle performance restituite dal tester e dal confronto delle stesse in ambedue le condizioni analizzate.
In queste condizioni, il simulatore opera dunque una Predizione del profilo d'espressione.

• **Subnetwork Simulation**
Questa tipologia di simulazione è stata condotta utilizzando come sorgente di estrazione della rete analizzata, il file .gpickle contenente la rete totale, ottenuta dall'unione di tutti i Simple-Paths estratti. La rete è stata dunque tagliata, di modo tale da estrarne un sottografo contenente *TLK2*, i geni di interesse e tutti i geni di congiunzione tra questi, ricavati a partire dai simple-paths, nonché tutte le interazioni presenti tra essi.
Il sottografo così estratto risulta essere dunque composto da 13 geni: *LHX6, DYNLL1, TLK1, ASF1A, SPATA1, PAX6, DYNLL2, TLK2, IRF4, IRF7, PAX5, ASF1B, CABIN1.*
Ogni simulazione è stata condotta ancora una volta considerando *TLK2* sempre espresso, per la motivazione illustrata precedentemente, mentre gli altri geni sono stati posti in 4 configurazioni diverse:

- Best Configuration, ottenuta a partire dai risultati migliori elaborati in fase di Single Gene Status Validation;
- Worst Configuration, ottenuta complementando l'espressione ottenuta in fase di Single Gene Status Validation;
- All-One Configuration, ottenuta considerando tutti i geni espressi;
- All-Zero Configuration, ottenuta considerando tutti i geni inibiti.

Le configurazioni possibili risultano essere al massimo quattro, in assenza di un'eventuale sovrapposizione di risultati, presente nel caso in cui la migliore/la peggiore configurazione fossero corrispondenti ad una tra la *All-One/All-Zero* e viceversa, riducendo il numero di configurazioni effettive a 2.

La motivazione alla base di questa simulazione risiede nella ricerca di una convalida dell'effettiva ed eventuale co-espressione dei geni analizzati, sempre in virtù della mancanza di dati relativi all'espressione degli stessi. Infatti, senza questi ulteriori dati, non sarebbe possibile asserire l'esistenza (o l'eventuale assenza) di fenomeni di co-espressione agenti tra i geni analizzati, se non facendo nuovamente riferimento agli studi effettuati sulla letteratura. Questa simulazione si prefigge dunque lo scopo di identificare la *"Best Configuration"*, ossia lo stato di espressione dei geni analizzati, in presenza dell'effetto combinato degli stessi, e, quindi, informazioni relative alla loro co-espressione.
In queste condizioni, il simulatore opera una Validazione del profilo d'espressione.

Simulatore d'espressione con PPI monostato

Questo simulatore non ha richiesto processing preliminare finalizzato all'eliminazione di determinate interazioni, permettendo così di effettuare un'analisi multilivello.
Per questa configurazione sono state effettuate 4 tipologie di simulazioni:

- Simple Paths Simulation;
- Subnetwork Simulation;
- Single-Edge Deletion Simulation;
- Clustered-Edges Total Deletion Simulation;

La simulazione sui Simple-Paths è stata condotta in maniera analoga al caso precedente.

• Subnetwork Simulation
Questa simulazione si differenzia dal caso precedente, in termini di processing e riduzione della rete.
Infatti, nonostante il taglio della rete limitato ai 13 nodi illustrati in precedenza, il calcolo dei simple-paths, effettuato utilizzando come sorgente/destinazione i nodi di nostro interesse e come nodo secondario ogni altro nodo presente nel grafo, ha richiesto un'ulteriore riduzione, a causa dei costi computazionali troppo elevati (ogni simulazione della rete totale, portava, in questo modo, all'identificazione di svariati milioni di percorsi). Si è, quindi, stabilito di considerare solo i percorsi esistenti tra i diversi nodi di nostro interesse, per ogni combinazione, che risultavano essere così limitati a circa 800 migliaia di unità, processabili in un tempo ragionevole.
Ciò ha inoltre permesso di avere gli stessi percorsi sia entranti che uscenti, riducendo ulteriormente il costo computazionale dovuto all'elaborazione delle liste contenenti i percorsi, che è stata elaborata una sola volta e poi semplicemente duplicata.
La simulazione si è poi svolta analogamente a quanto già visto.

• Single-edge deletion Simulation
Questa tipologia di simulazione è stata condotta utilizzando nuovamente come sorgente di estrazione della rete analizzata, il file .gpickle contenente la rete totale, opportunamente tagliata e processata come visto in precedenza. Anche il calcolo dei percorsi è stato limitato. Va segnalato che ogni rete è stata simulata nella configurazione di espressione riportante le performance migliori in fase di Validazione del profilo d'espressione.

É stata inoltre utilizzata la routine di identificazione degli archi uscenti tra TLK2, i quali rappresentano gli effetti regolatori o di interazione messi in atto da TLK2; va segnalato che tutti gli archi sono risultati essere appartenenti alla categoria di interazione proteica. Ogni arco uscente da TLK2 e rappresentante una funzionalità dello stesso e delle proteine da esso codificate, è stato così identificato e salvato in un apposito file, contenente la lista totale dei 12 archi di nostro interesse.
Successivamente, gli archi sono stati rimossi singolarmente, testando la rete e valutando le modifiche subite dalla stessa ad ogni rimozione.
La motivazione alla base di questa simulazione risiede nella volontà di poter quantificare e qualificare l'effetto di una eventuale mutazione in TLK2 e, nel caso di nostro interesse, nelle interazioni proteiche che lo coinvolgono. Risulta infatti ampiamente noto, tramite studi di letteratura, che sono molteplici i fattori derivati da una mutazione, che possono influenzare e impedire un'interazione proteica, come dei cambiamenti nel sito di fosforilazione, la sostituzione con amminoacidi con proprietà chimico-fisiche distanti da quelli originali, la vicinanza della mutazione rispetto all'interfaccia di interazione, la variazione di stabilità del complesso, modifiche strutturali che possono riguardare la conformazione "ripiegata" delle proteine. Questa simulazione si prefigge dunque lo scopo di simulare una mutazione agente sulle PPI normalmente riguardanti TLK2, qualificandone gli effetti e permettendo così una validazione delle altre proteine appartenenti ai diversi complessi, mediante valutazione delle performance ottenute.
In queste condizioni, il simulatore opera una Validazione delle proteine correlate alla mutazione.

• **Clustered edges deletion Simulation** Questa tipologia di simulazione è stata svolta con un set-up perfettamente analogo a quanto visto in fase di Single-Edge Deletion Simulation.
In questo caso, però, si è proceduto alla rimozione totale degli archi appartenenti a due Cluster:

- Bad-Performance Cluster, comprendente quegli archi la cui rimozione, in fase di validazione proteica, ha riportato un peggioramento delle performance generali della rete analizzata;
- Good-Performance Cluster, comprendente quegli archi la cui rimozione, in fase di validazione proteica, ha causato un miglioramento delle performance generali della rete analizzata;

Come detto, sono state effettuate 2 simulazioni distinte, finalizzate alla rimozione totale di tutti gli archi inclusi nei due distinti Cluster. La motivazione alla base di questa simulazione risiede nella volontà di identificare l'effetto combinato, dovuto alla presenza contemporanea delle varie mutazioni riguardanti le interazioni proteiche incluse nei diversi cluster. In letteratura, infatti, è risaputo che diverse interazioni presenti in maniera contemporanea, potrebbero avere due tipi di effetti:

- Positivo, nella quale la presenza combinata di più mutazioni non aggiunge danno effettivo rispetto alla presenza delle singole mutazioni (come nel caso di 2 entità che cooperano per il raggiungimento di uno scopo unico, impedito già con l'eliminazione dell'interazione riguardante una delle due proteine), oppure la presenza combinata di diverse mutazioni implica una qualche soppressione a cascata degli effetti delle stesse;

– Negativo, nel quale la presenza combinata di più mutazioni assume una connotazione deleteria rispetto alla presenza delle singole mutazioni.

In questo modo è stato possibile indagare ulteriormente l'effetto delle mutazioni presenti nei 2 cluster e soprattutto validarne ulteriormente l'appartenenza ad uno di essi, tramite valutazione delle performance ottenute.
In queste condizioni il tester opera una Validazione dei Clusters proteici.

Simulatore d'espressione con PPI monostato valutata

Anche questo simulatore non ha richiesto processing preliminare finalizzato all'eliminazione di determinate interazioni.
Per questa configurazione sono state effettuate 4 tipologie di simulazioni:

- Simple Paths Simulation;
- Subnetwork Simulation;
- Single-Edge Deletion Simulation;
- Clustered-Edges Total Deletion Simulation;
- Clustered-Edges 2-Combinations Deletion Simulation;

Tutte le simulazioni, comuni al caso precedente, sono state svolte in maniera analoga.

• Clustered edges 2-combinations deletion Simulation
Questa tipologia di simulazione è stata svolta con un set-up perfettamente analogo a quanto visto in fase di Single-Edge Deletion Simulation e Clustered-Edge Total Deletion Simulation.
In questo caso, però, si è proceduto alla rimozione separata degli archi appartenenti ai due Cluster, per combinazioni di elementi presi 2 a 2.
La motivazione alla base di questa simulazione risiede nella volontà di identificare in maniera ancora più specifica l'effetto combinato, dovuto alla presenza contemporanea di 2 mutazioni riguardanti le interazioni proteiche incluse nei diversi cluster. In questo modo è stato possibile indagare e qualificare in maniera più dettagliata l'effetto combinato delle mutazioni presenti nei 2 cluster.
In queste condizioni il tester opera una Definizione dell'effetto combinato di mutazioni doppie.

Simulatore d'espressione con PPI bistato valutata

Anche questo ulteriore simulatore non ha richiesto processing preliminare finalizzato all'eliminazione di determinate interazioni.
Per questa configurazione sono state effettuate 4 tipologie di simulazioni:

- Simple Paths Simulation;
- Subnetwork Simulation;
- Single-Edge Deletion Simulation;
- Clustered-Edges Total Deletion Simulation;
- Clustered-Edges 2-Combinations Deletion Simulation;

Tutti le simulazioni sono state svolte in maniera analoga al caso precedente.

Capitolo 13

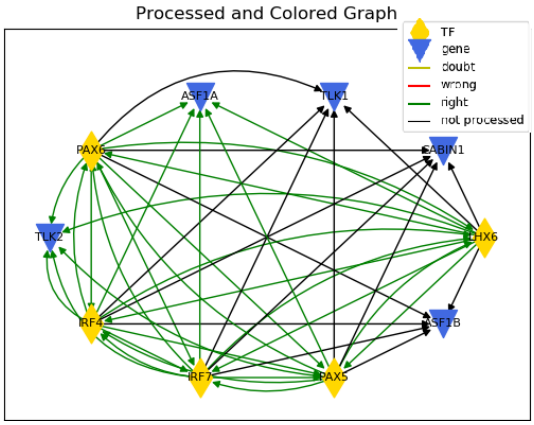
Risultati

Qui di seguito sono riportati i risultati relativi a tutte le simulazioni effettuate. Ogni risultato ottenuto è illustrato mediante lo stesso grafo analizzato opportunamente processato e colorato in base allo stato predetto (rombo=TF,triangolo=gene,arco giallo=dubbio, arco rosso=sbagliato, arco verde=corretto, arco nero=**non processato**), e due rappresentazione grafiche del Report, in particolare un andamento dei valori percentuali e un istogramma contenente la media dei valori percentuali.

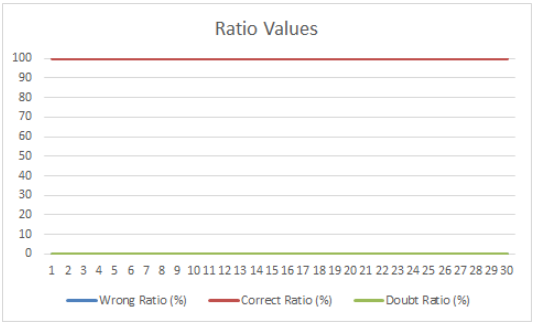
Simulatore d'espressione della GRN

• **Predizione del profilo d'espressione**
In Figura 13.1 sono mostrati i risultati relativi alla validazione effettuata considerando il subnetwork ottenuto dall'estrazione di tutti i Simple Paths esistenti tra *TLK2* e *ASF1A*, considerando*ASF1A* espresso.
Come si può notare tutti gli archi processati risultano essere corretti.
Indagando ulteriormente i percorsi identificati è stato possibile scoprire che, successivamente all'eliminazione degli archi di tipo PPI, non sono rimasti percorsi che collegano i 2 geni analizzati, vedi Figura 13.1a. Risultati analoghi sono stati ottenuti sia modificando lo stato di espressione di *ASF1A* in repressione, che analizzando le interazioni con *ASF1B* e *TLK1*, nei due diversi stati di espressione.

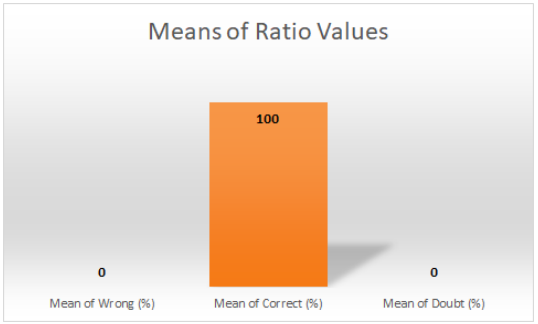
Si potrebbe affermare che gli archi di tipo PPI risultano essere fondamentali per la predizione del *Best Status* di ogni gene, operazione che viene applicata alle reti in nostro possesso, perciò la loro eliminazione rende questo tentativo di predizione fallimentare.



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



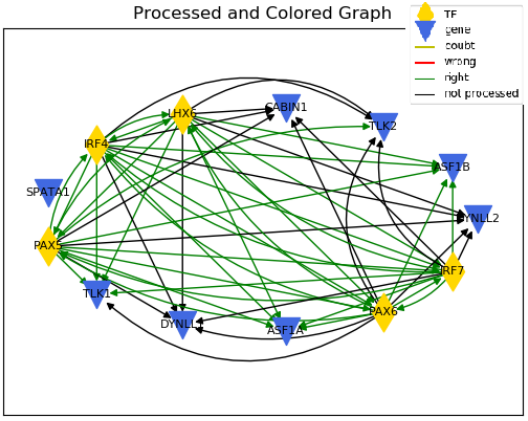
(c) *Media dei Valori Percentuali*

Figura 13.1: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione della GRN, con ASF1A attivo)*

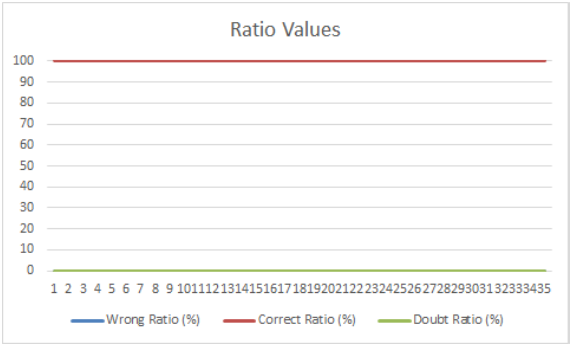
- **Validazione del profilo d'espressione**

È stato analizzato il sottografo completo contenente i 4 geni di nostro interesse, ai quali sono stati assegnati arbitrariamente, data la mancanza di dati provenienti dalla *predizione del profilo d'espressione*, i due diversi stati di espressione.
In Figura 13.2 sono mostrati i risultati relativi alla validazione effettuata sulla suddetta rete di co-espressione, considerando tutti e 4 i geni attivi. Ancora una volta è possibile notare come tutti gli archi processati sono corretti, in quanto l'eliminazione degli archi di interazione proteica ha fatto sì che non esistessero percorsi di collegamento fra i geni di nostro interesse, vedi grafo riportato in Figura 13.2.

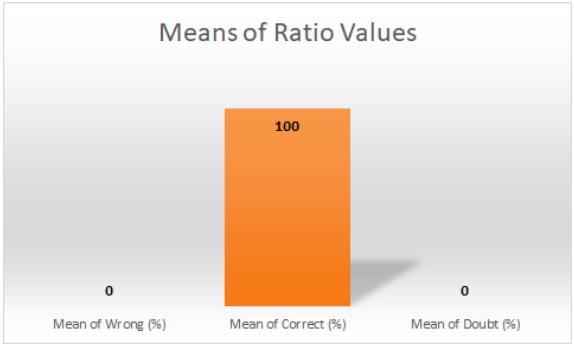
Ancora una volta gli archi di PPI risultano essere fondamentali per la predizione della migliore configurazione della Rete di Co-espressione genica analizzata, rendendo questa ulteriore validazione fallimentare. Non è stato possibile proseguire con le simulazioni poiché tutti gli archi uscenti da *TLK2* risultavano essere, come già detto, di interazione proteica, rendendo quindi l'emulazione della mutazione impossibile. Possiamo pertanto affermare che, sebbene questo simulatore sia concettualmente corretto, in quanto separa il problema della regolazione genica da quello dell'interazione proteica, non risulta essere adatto al problema considerato. Tutto ciò è dovuto sia all'origine dei dati di partenza, che sono stati ottenuti analizzando l'interattoma, ma anche alla rete stessa analizzata, che rende necessaria l'implementazione degli archi di interazione proteica al fine di analizzare le cause e gli effetti di diversi percorsi agenti tra le entità di nostro interesse.



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



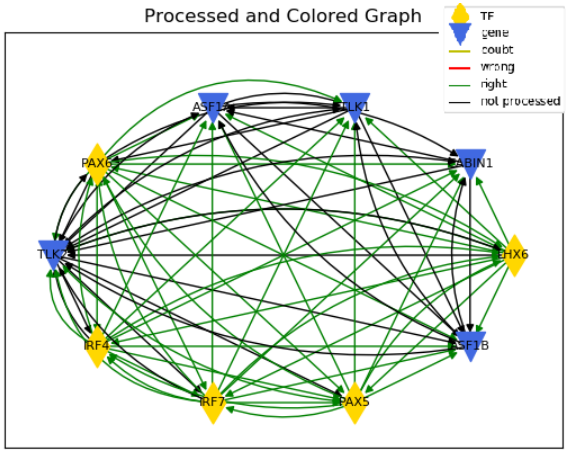
(c) *Media dei Valori Percentuali*

Figura 13.2: *Risultati della validazione del profilo d'espressione (Simulatore d'espressione della GRN, All-One Configuration)*

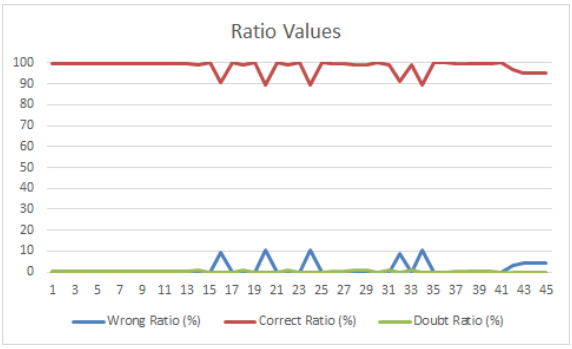
Simulatore d'espressione con PPI monostato

• **Predizione del profilo d'espressione**

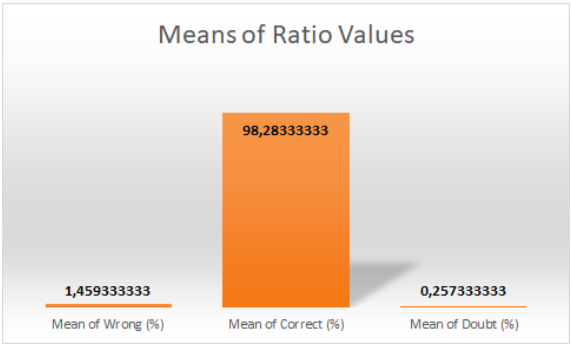
In Figura 13.3 e in Figura 13.4 sono mostrati i risultati relativi alla validazione effettuata considerando la sottorete relativa a *TLK2* e *ASF1A*, forzando rispettivamente lo stato di *ASF1A* come acceso e spento. I grafi riportano, in ambedue i casi, gli archi come tutti corretti, poiché il numero di archi sbagliati in proporzione risulta essere sempre troppo basso. Effettuando in seguito un confronto tra le varie medie dei Valori Percentuali, è possibile vedere che nel primo caso, ossia con l'accensione di *ASF1A*, si assiste ad un innalzamento della media degli archi corretti ed un abbassamento di quella degli archi dubbi, a patto di un leggero peggioramento negli archi sbagliati. Per questo motivo è stato scelto come *Best Status* di *ASF1A* l'attivazione.



(a) *Grafo processato e colorato secondo gli stati predetti*

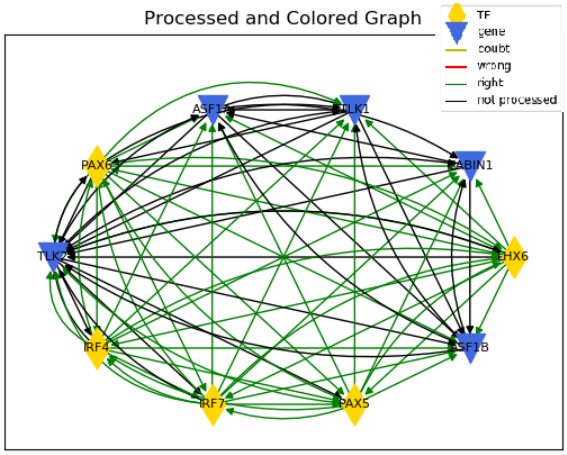


(b) *Andamento dei Valori Percentuali*

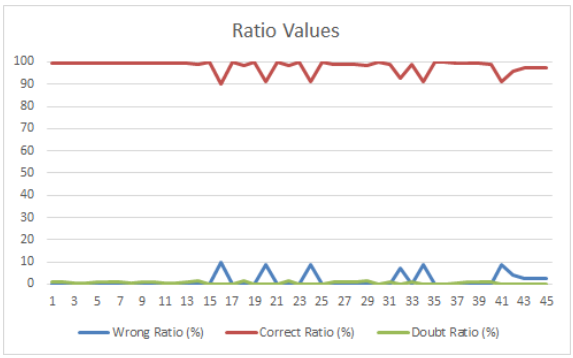


(c) *Media dei Valori Percentuali*

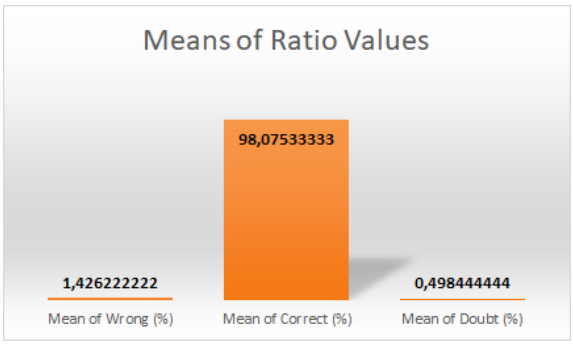
Figura 13.3: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato, con ASF1A attivo)*



(a) *Grafo processato e colorato secondo gli stati predetti*



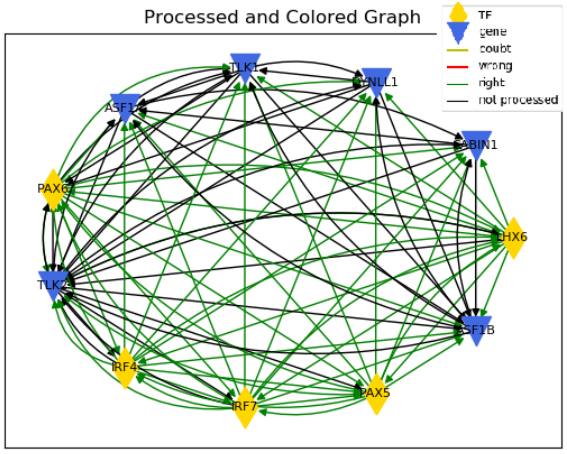
(b) *Andamento dei Valori Percentuali*



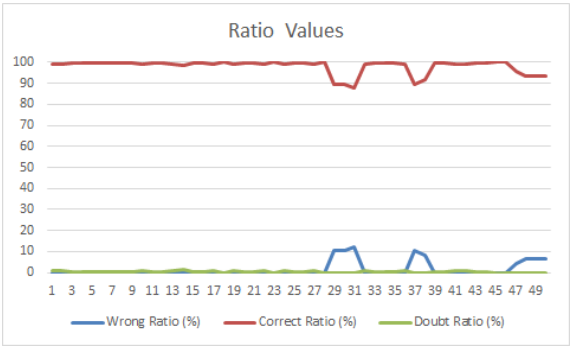
(c) *Media dei Valori Percentuali*

Figura 13.4: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato, con ASF1A inibito)*

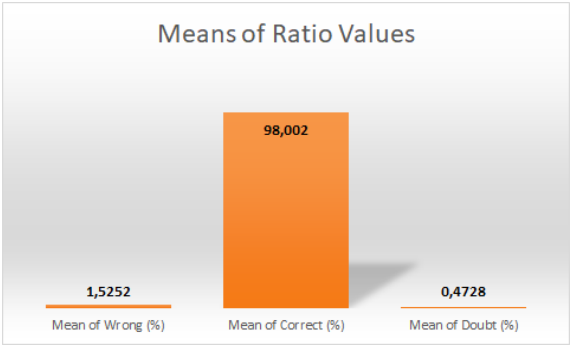
Valutazioni analoghe sono state effettuate con *ASF1B* e *TLK1*, per i quali il best status è risultato essere sempre l'attivazione. I risultati per questa configurazione, per entrambi i geni, sono mostrati in Figura 13.5 e 13.6.



(a) *Grafo processato e colorato secondo gli stati predetti*

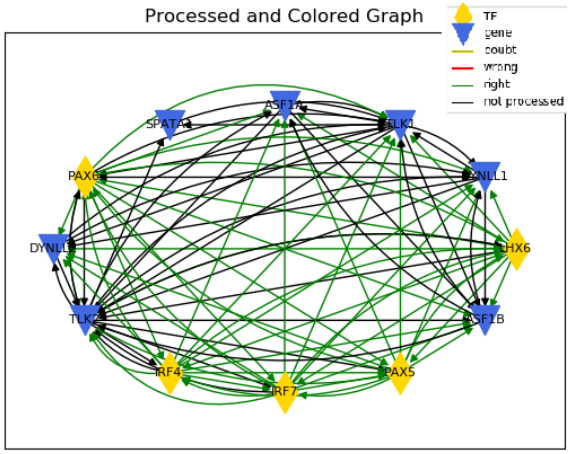


(b) *Andamento dei Valori Percentuali*

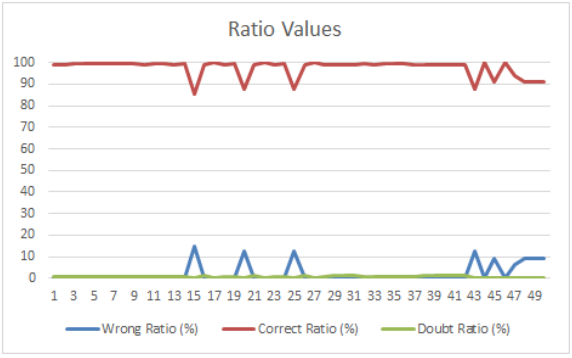


(c) *Media dei Valori Percentuali*

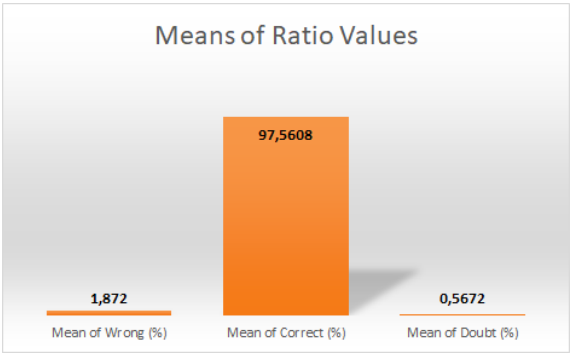
Figura 13.5: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato, con ASF1B attivo)*



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



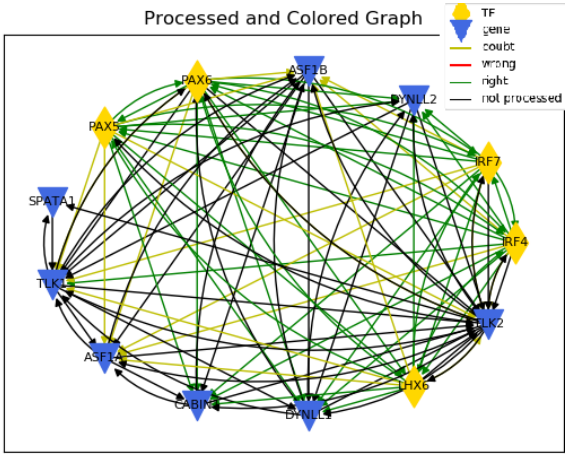
(c) *Media dei Valori Percentuali*

Figura 13.6: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato, con TLK1 attivo)*

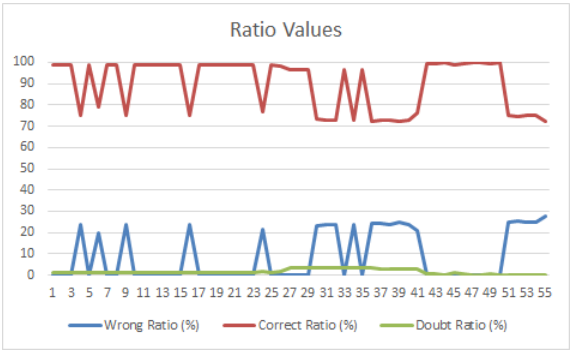
I risultati ottenuti mostrano di essere allineati con gli studi riportati nel paragrafo introduttivo presente nella Sezione IV, in quanto tutti e 3 i geni risultano essere in condizioni di normalità quasi sempre espressi, mostrando caratteristiche di co-espressione con *TLK2*.
Il simulatore dimostra così un certo grado di coerenza con la letteratura, che ne valida l'efficacia.

- **Validazione del profilo d'espressione**

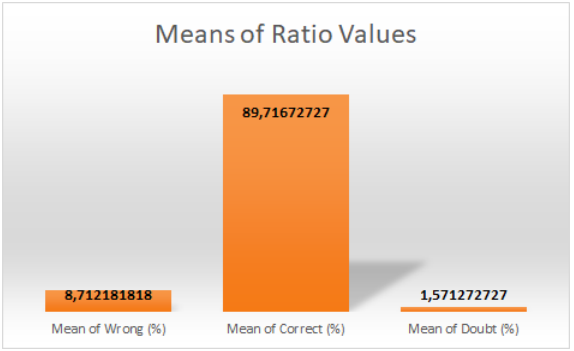
È stato analizzato il sottografo completo, opportunamente processato, contenente i 4 geni di nostro interesse, ai quali sono stati assegnate 2 configurazioni di espressione e mantenendo *TLK2* sempre attivo: tutti e 3 i geni attivi (*All-One/Best*) e la configurazione complementare (*All-Zero/Worst*).
Come detto nel Capitolo 12, queste configurazioni sono state ottenute analizzando i risultati prodotti dalla *predizione del profilo d'espressione*.
In Figura 13.7 sono mostrati i risultati relativi alla validazione effettuata sulla suddetta rete di co-espressione, considerando tutti e i geni attivi, mentre in Figura 13.8 sono mostrati i risultati ottenuti considerando la configurazione complementare.
Valutando gli andamenti e le medie dei valori percentuali, è possibile notare un netto aumento di performance apportato dal forzamento della configurazione All-One, rispetto alla sua speculare, che mostra ancora una volta l'allineamento e la congruenza con le informazioni presenti in letteratura, riportate nel paragrafo introduttivo della Sezione IV, che evidenziano una la co-espressione e la cooperazione esistente tra i 4 geni di nostro interesse.



(a) *Grafo processato e colorato secondo gli stati predetti*

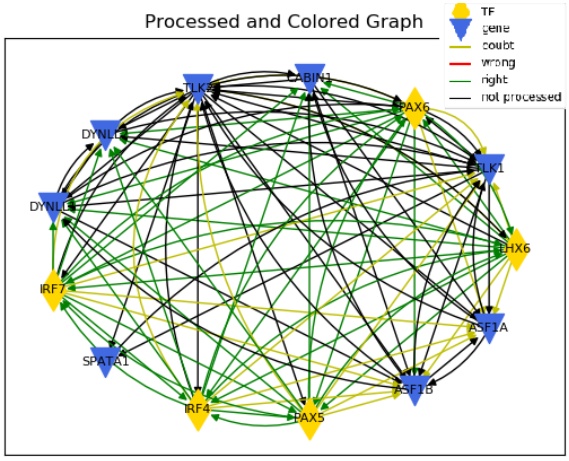


(b) *Andamento dei Valori Percentuali*

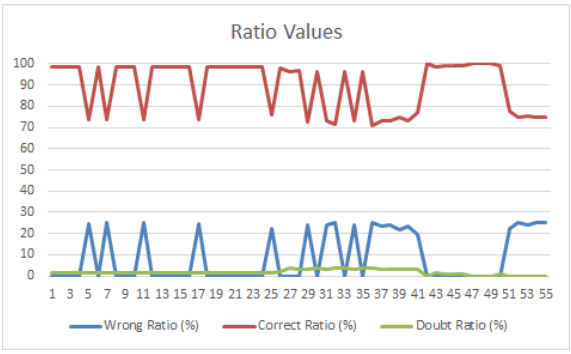


(c) *Media dei Valori Percentuali*

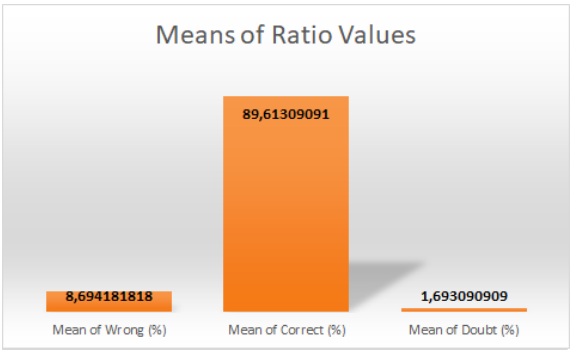
Figura 13.7: *Risultati della validazione del profilo d'espressione (Simulatore d'espressione con PPI monostato, All-One/Best Configuration)*



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



(c) *Media dei Valori Percentuali*

Figura 13.8: *Risultati della validazione del profilo d'espressione (Simulatore d'espressione con PPI monostato, All-Zero/Worst Configuration)*

Tutte le simulazioni successive sono state dunque condotte utilizzando come profilo d'espressione il migliore ricavato in questa fase, ossia la *All-One Configuration*.

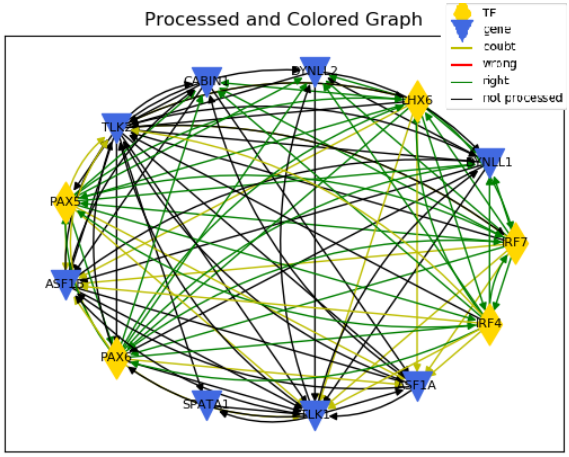
- **Validazione delle proteine correlate alla mutazione**

Tutti gli archi uscenti da *TLK2*, rappresentanti gli effetti dovuti alla sua espressione, sono stati identificati e singolarmente eliminati al fine di simulare l'effetto di una mutazione, come detto nel capitolo 12. Infatti, i dati forniti dai ricercatori di scienze mediche-genetiche dell'Università di Torino, trattati nel paragrafo introduttivo della Sezione IV, riguardano 2 varianti missenso di *TLK2*, ossia mutazioni che hanno come effetto la codifica di un amminoacido diverso da quello originario, con un conseguente effetto sull'interazione con gli altri geni.
Tutti gli archi identificati, come già spiegato precedentemente, sono risultati essere archi di interazione proteica, ossia non collegati direttamente ad interazioni genetiche, ma ad un prodotto degli stessi. La mutazione simulata, di conseguenza, avrebbe effetto sull'interazione proteica considerata.
Sono diversi gli effetti che una mutazione può avere sull'interazione proteica e si distinguono in base al tipo di interazione, che può essere funzionale, fisica, diretta, indiretta, o in base alla vicinanza dell'interazione rispetto al binding site, o anche alle differenze dei residui amminoacidici sostituiti.
Le performance risultanti da ogni eliminazione sono state valutate. La prima caratteristica identificata, come riportato nel Capitolo 12, è stata l'appartenenza delle mutazioni a due Cluster distinti:

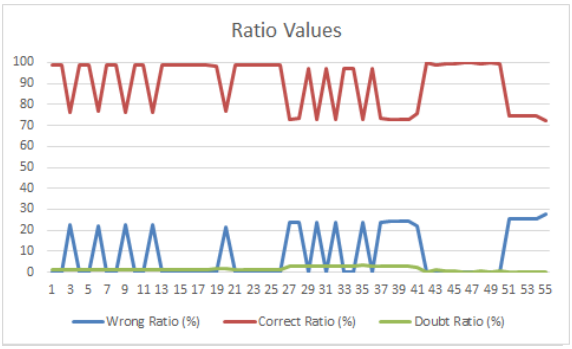
  - *Bad-Performance Cluster*, comprendente tutti gli archi la cui rimozione provoca un peggioramento delle performance;

  - *Good-Performance Cluster*, comprendente gli archi la cui rimozione provoca un miglioramento delle performance.

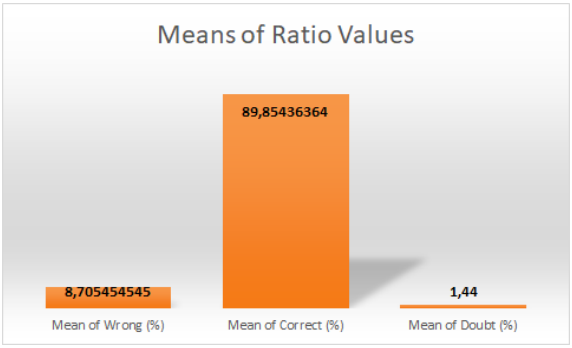
In Figura 13.9 sono riportati i risultati relativi alla mutazione appartenente al *Bad-Performance Cluster* che restituisce il decremento maggiore di performance, ossia la delezione dell'interazione *TLK2-PAX6*; mentre in Figura 13.10 sono riportati quelli relativi alla mutazione appartenente all'altro cluster, che restituisce l'incremento maggiore di performance, ossia la delezione dell'interazione *TLK2-IRF4*.



(a) *Grafo processato e colorato secondo gli stati predetti*

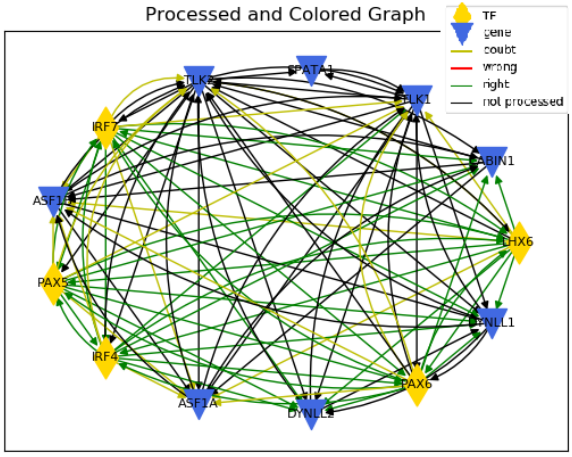


(b) *Andamento dei Valori Percentuali*

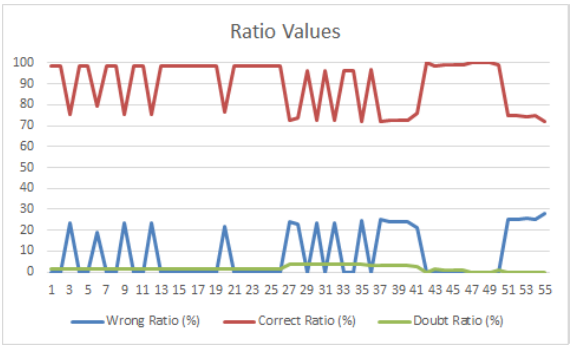


(c) *Media dei Valori Percentuali*

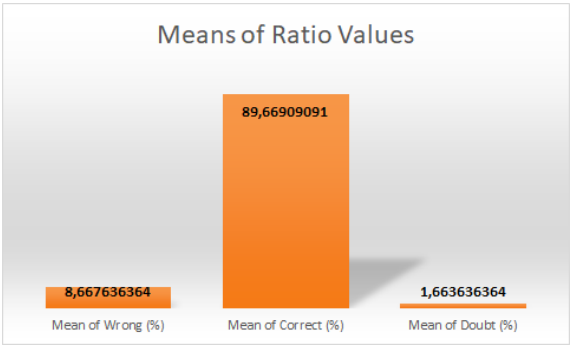
Figura 13.9: *Risultati della validazione delle proteine associate alla mutazione (Simulatore d'espressione con PPI monostato, All-One/Best Configuration, delezione PPI TLK2-IRF4)*



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



(c) *Media dei Valori Percentuali*

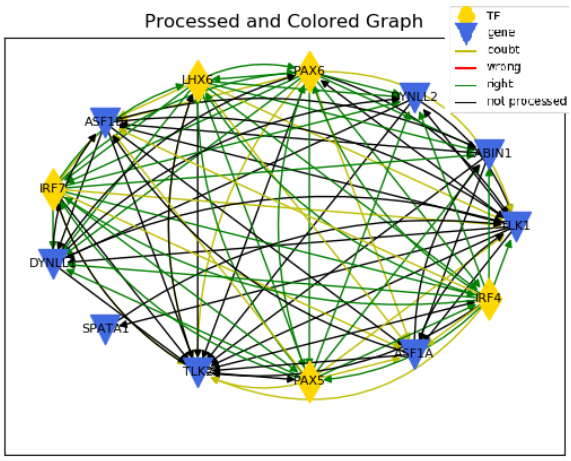
Figura 13.10: *Risultati della validazione delle proteine associate alla mutazione (Simulatore d'espressione con PPI monostato, All-One/Best Configuration, delezione PPI TLK2-PAX6)*

Analizzando le interazioni tramite *TheRingDB*, è risultato che quasi tutte le PPI analizzate sono dirette e di associazione fisica. Mutazioni agenti su interazioni di questo tipo potrebbero modificare l'energia di legame, in maniera tanto più deleteria quanto la sostituzione amminoacidica riguardi amminoacidi prossimi al binding site e con proprietà fisico-chimiche molto differenti da quelli sostituiti.

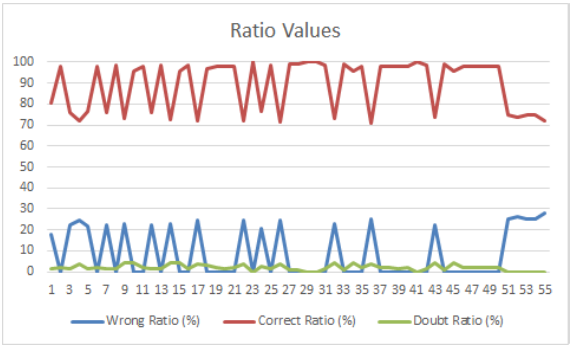
Per quanto riguarda, invece, le mutazioni lontane dall'interfaccia, esse potrebbero provocare un errore di folding.
Dalla letteratura inoltre, come riportato nel paragrafo introduttivo della Sezione IV, risulta che le proteine prodotte dai geni di nostro interesse partecipano in percorsi segnalatori, attraverso fosforilazione; una conseguenza di una mutazione, infatti, potrebbe coinvolgere proprio il sito di fosforilazione, causando una perdita di molte proprietà funzionali.
Per quanto riguarda il problema analizzato, è stato deciso di focalizzarsi sulle proteine che migliorano le performance della rete, ossia appartenenti al *Good-Performance Cluster*. Questo perché, migliorando le performance, validerebbero ancora di più il profilo d'espressione di partenza, indicando, dunque, una sovraespressione dei geni di nostro interesse. Infatti, come detto nel paragrafo introduttivo della Sezione IV, l'aumento d'interazione proteica apportato dalle 2 mutazioni missense, verificatosi in tutti i geni di nostro interesse, è stato correlato ad una sovraespressione degli stessi.
I geni che producono proteine che hanno un'interazione proteica con *TLK2* di questo tipo risultano essere 4: *IRF7, IRF4, PAX5, LHX6*.
Sapendo che l'interazione proteica riguarda *TLK2*, che risulta appartenere ai geni con espressione nota, possiamo dedurre che per via della propagazione dello stato applicata da questa tipologia di simulatore anche l'altro gene dovrebbe essere, in maniera indiretta, espresso in condizioni di normalità. In virtù di ciò, si potrebbe ipotizzare che la mancata interazione tra le due entità proteiche possa in un qualche modo impedire la generazione di prodotti e complessi che favoriscono l'espressione del gene secondario coinvolto nell'interazione, ad esempio nel caso di proteine agenti in maniera combinata sui geni, come *Co-Transcription Factor*.
Un'altra ipotesi possibile è che la mancata interazione tra le due entità proteiche possa avere direttamente un effetto sulle sorgenti di nostro interesse, impedendone l'inibizione quando necessaria.
Inoltre, la mutazione delle stesse proteine, prescindendo dall'interazione tra esse, potrebbe avere effetto sull'espressione dei geni, ad esempio nel caso di proteine che agiscono direttamente sui geni, come *Transcription Factor*.
In ultima battuta, va segnalato che ogni interazione proteica è il riflesso di un processo adattativo ed evolutivo; in virtù di ciò, i cambiamenti che si verificano, magari sotto effetto di una mutazione, in una delle proteine coinvolte in un'interazione, potrebbero essere compensati da cambiamenti nell'altra. Dunque, una mutazione riguardante una delle due entità coinvolte, potrebbe anche riflettersi a cascata sulle altre entità che interagiscono su essa, provocando eventualmente una degradazione maggiore della stabilità dei meccanismi funzionali presenti nell'organismo.
Questo tipo di validazione ha quindi permesso l'identificazione di un pool di 4 proteine, nonché delle relative interazioni proteiche da segnalare ai biologi, in modo tale da studiarli in maniera approfondita, al fine di poter trovare correlazioni con lo sviluppo di autismo.
Va però segnalato che non è stato possibile stabilire, in base ai dati disponibili, se le mutazioni ricercate in *TLK2* hanno effetto su più relazioni, e che una mutazione potrebbe anche rafforzare l'interazione proteica, piuttosto che eliderla. Non è stato inoltre possibile discriminare le due varianti missense.

- **Validazione dei Clusters proteici**

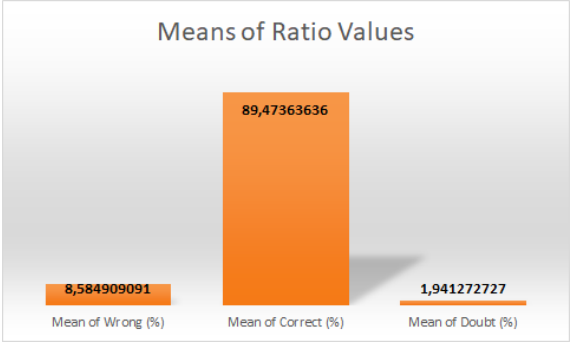
In questo caso, tutte le interazioni appartenenti ai due Cluster sono state eliminate dalla rete in maniera sincrona, ottenendo così le performance della rete dovute all'eliminazione di tutte le interazioni proteiche nel *Bad-Performance* e nel *Good-Performance Clusters*, mostrate rispettivamente in Figura 13.11 e in Figura 13.12.



(a) *Grafo processato e colorato secondo gli stati predetti*

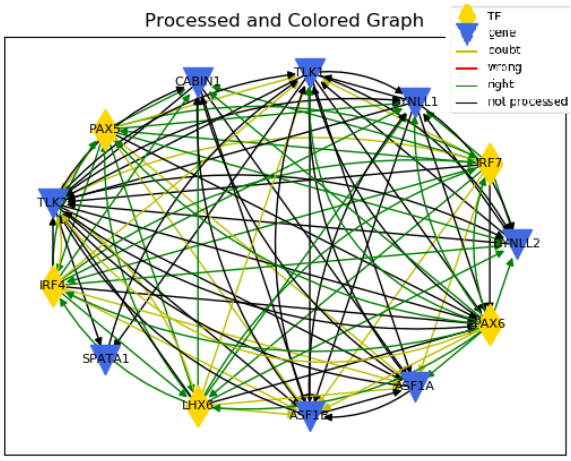


(b) *Andamento dei Valori Percentuali*

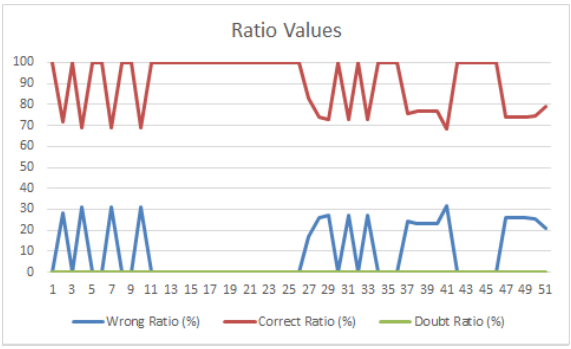


(c) *Media dei Valori Percentuali*

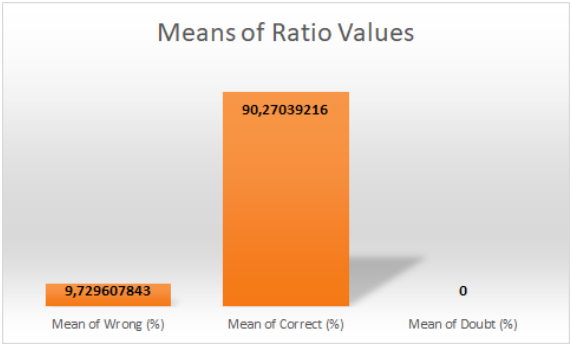
Figura 13.11: *Risultati della validazione dei clusters proteici (Simulatore d'espressione con PPI monostato, All-One/Best Configuration, delezione Bad-Performance Cluster)*



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



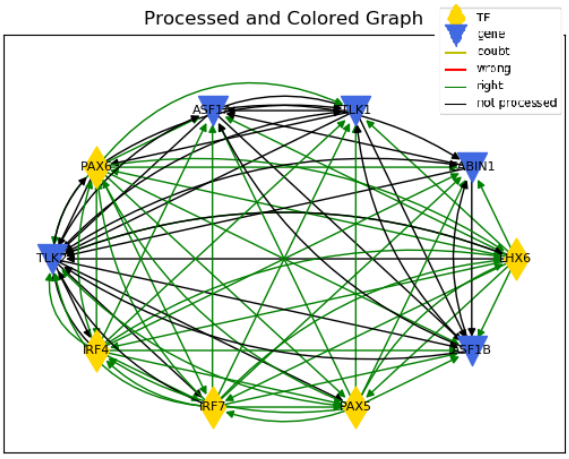
(c) *Media dei Valori Percentuali*

Figura 13.12: *Risultati della validazione dei clusters proteici (Simulatore d'espressione con PPI monostato, All-One/Best Configuration, delezione Good-Performance Cluster)*

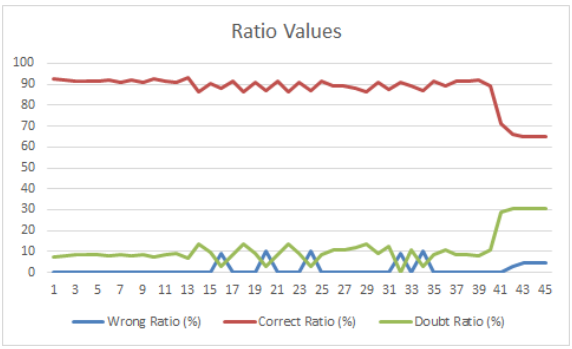
È possibile notare come le performance diminuiscano notevolmente nel caso del primo Cluster, in accordo con quando ipotizzato in fase di delezione degli archi singoli. Nel secondo Cluster le performance aumentano notevolmente, provocando inoltre un'eliminazione di tutti i percorsi dubbi.
Inoltre, per il *Good-Performance Cluster* varia anche il tempo di simulazione ed il numero di archi processati, come mostrato in Figura 13.12b. Ciò indica che i nodi analizzati sono degli Hubs, per via dell'alto coinvolgimento degli archi, contenuti nel suddetto Cluster, in percorsi che non vengono più identificati, a causa della loro rimozione.
Questa simulazione riesce a validare nuovamente il pool proteico e di interazioni identificato nella fase precedente.

Simulatore d'espressione con PPI monostato valutata

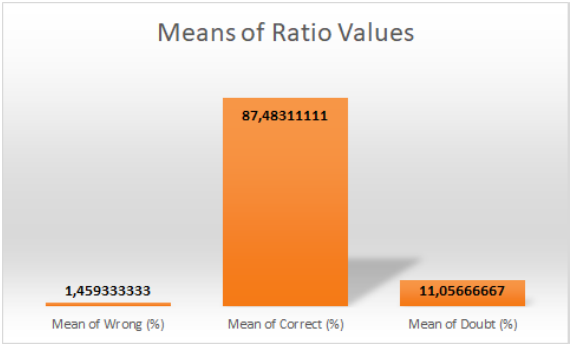
• Predizione del profilo d'espressione
Le simulazioni svolte hanno mostrato risultati analoghi al caso precedente.
In Figura 13.13 e 13.14 sono mostrati i risultati ottenuti simulando la sottorete relativa ai soli *TLK2 e ASF1A*, forzando rispettivamente l'accensione e lo spegnimento del secondo.
È possibile notare che la differenza tra le performance ottenute dal forzamento dei due stati risulta essere molto più marcata a causa dell'integrazione della valutazione di correttezza dovuta all'elaborazione degli archi di PPI, la quale in caso di errore ha un risvolto immediato nel numero di archi dubbi identificati nei percorsi regolatori. Ciò nonostante, il Best Status di *ASF1A* risulta essere quello di attivazione.



(a) *Grafo processato e colorato secondo gli stati predetti*

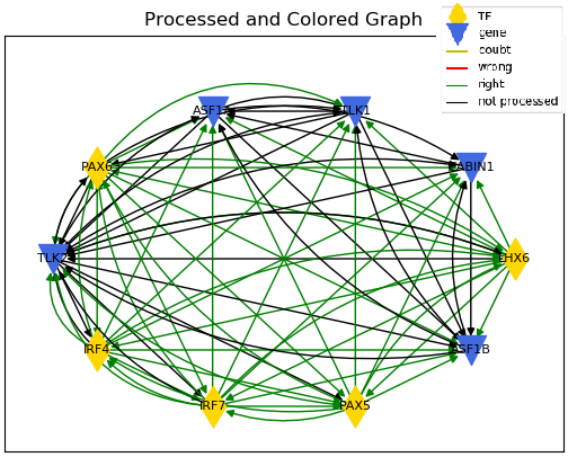


(b) *Andamento dei Valori Percentuali*

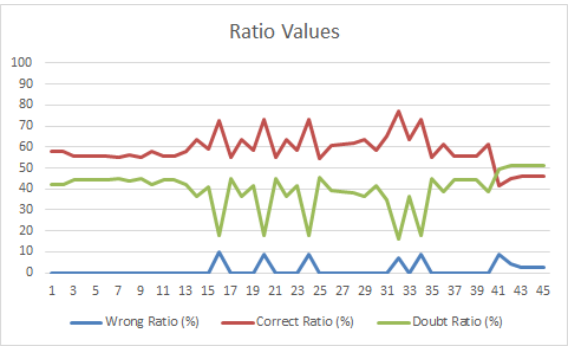


(c) *Media dei Valori Percentuali*

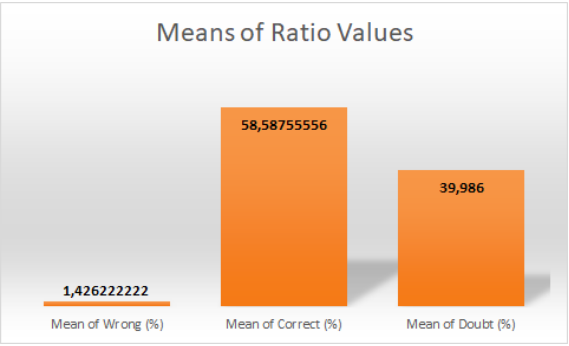
Figura 13.13: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato valutata, con ASF1A attivo)*



(a) *Grafo processato e colorato secondo gli stati predetti*



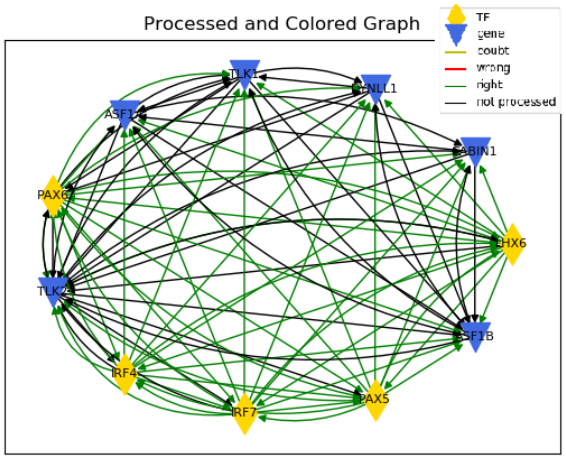
(b) *Andamento dei Valori Percentuali*



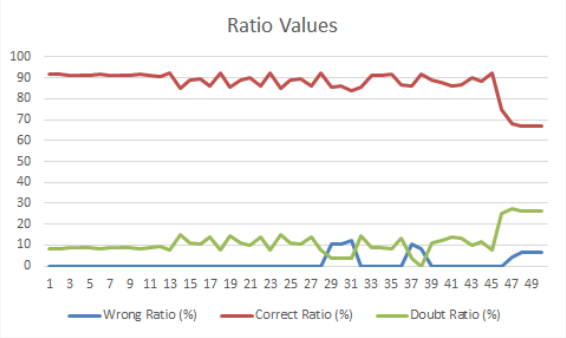
(c) *Media dei Valori Percentuali*

Figura 13.14: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato valutata, con ASF1A inibito)*

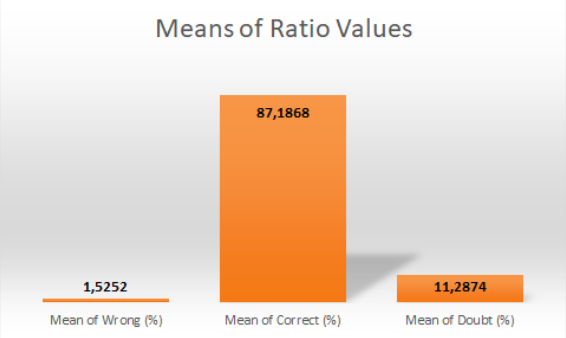
Ancora una volta, il *Best Status* dei 3 geni è risultato essere di attivazione. Valutazioni analoghe sono state fatte con *ASF1B e TLK1*, per i quali il best status è risultato essere sempre l'attivazione; i risultati per questa configurazione, per entrambi i geni, sono mostrati in Figura 13.15 e 13.16.



(a) *Grafo processato e colorato secondo gli stati predetti*

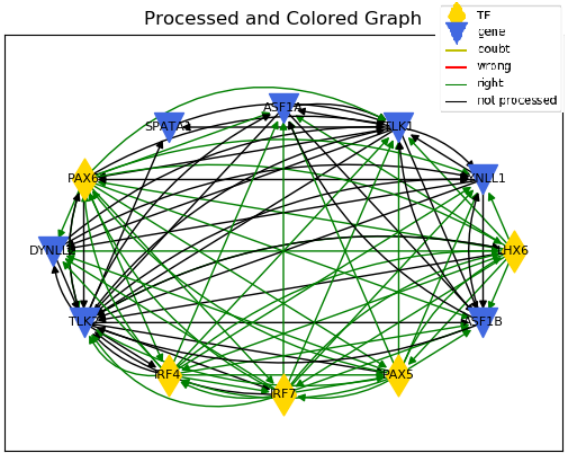


(b) *Andamento dei Valori Percentuali*

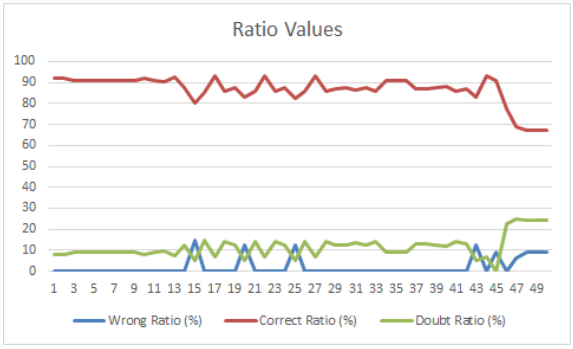


(c) *Media dei Valori Percentuali*

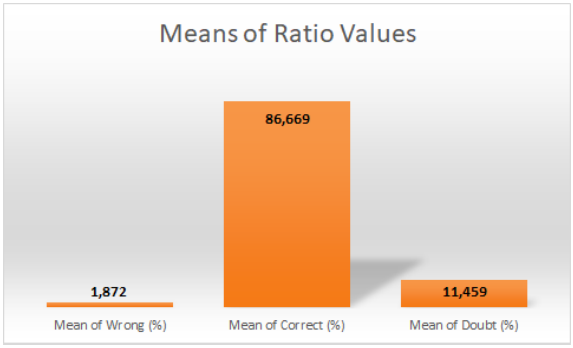
Figura 13.15: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato valutata, con ASF1B attivo)*



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



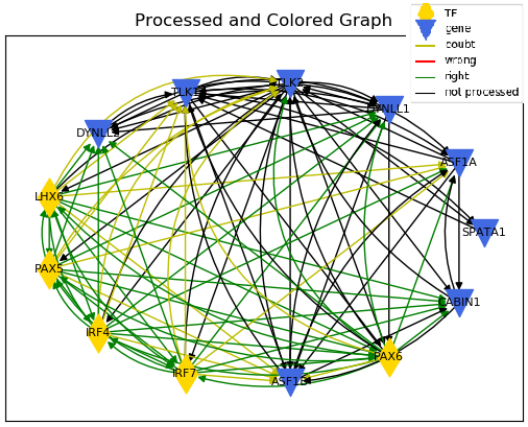
(c) *Media dei Valori Percentuali*

Figura 13.16: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI monostato valutata, con TLK1 attivo)*

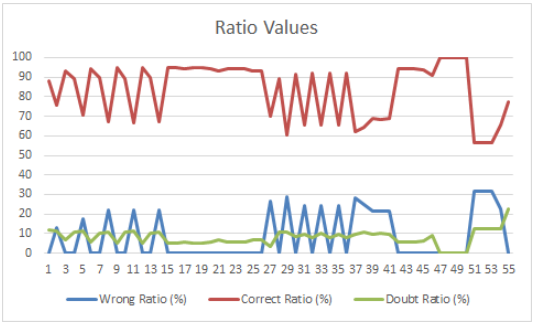
Anche in questo caso, il simulatore risulta essere allineato ai risultati forniti in letteratura, evidenziando, però, differenze molto più marcate.

- **Validazione del profilo d'espressione**

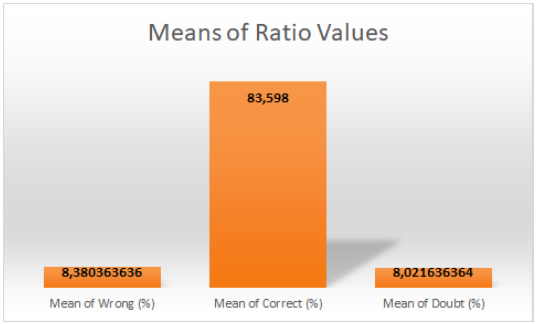
È stato dunque analizzato il sottografo completo, opportunamente processato, contenente i 4 geni di nostro interesse, ai quali sono stati assegnate le due configurazioni di espressione: *All-One/Best* e la configurazione complementare, *All-Zero/Worst*.
Anche in questo caso, è stato possibile notare un netto aumento di performance apportato dal forzamento della configurazione *All-One*, i cui risultati sono mostrati in Figura 13.17, rispetto alla sua speculare. I risultati sono nuovamente congruenti a quanto riportato in letteratura.



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



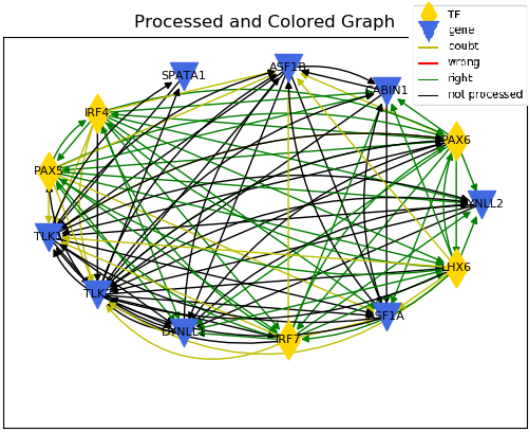
(c) *Media dei Valori Percentuali*

Figura 13.17: *Risultati della validazione del profilo d'espressione (Simulatore d'espressione con PPI monostato valutata, All-One/Best Configuration)*

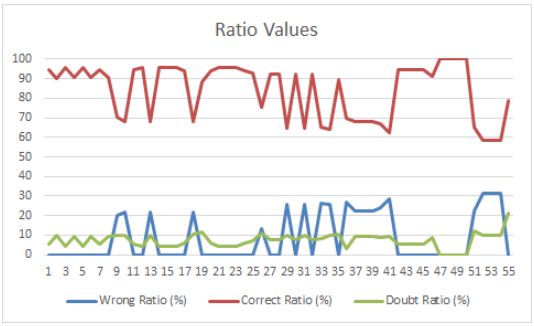
Tutte le simulazioni successive sono state dunque condotte utilizzando come profilo d'espressione la *All-One Configuration*.

- **Validazione delle proteine correlate alla mutazione**

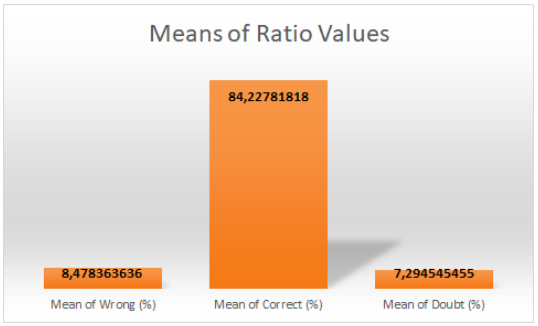
Anche in questa simulazione, sono stati identificati i due Cluster riportati precedentemente. Il Cluster di nostro interesse, ossia contenente gli archi la cui rimozione apporta un miglioramento di performance, risulta essere sempre composto dai 4 geni identificati con la prima versione del simulatore: *IRF4, IRF7, LHX6, PAX5*.
In Figura 13.18 sono riportate le performance dovute all'eliminazione dell'arco di PPI *TLK2-IRF4*, che restituisce l'incremento di performance maggiori.



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



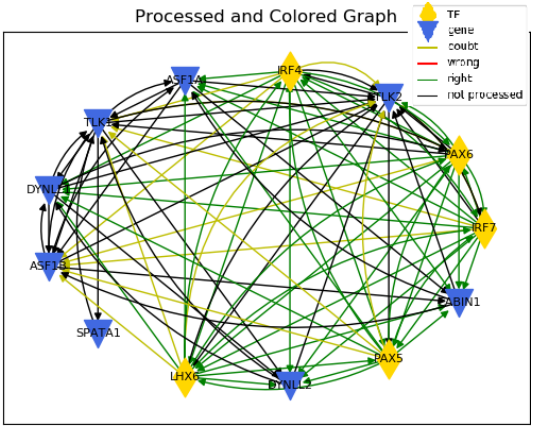
(c) *Media dei Valori Percentuali*

Figura 13.18: *Risultati della validazione delle proteine correlate alla mutazione (Simulatore d'espressione con PPI monostato valutata, All-One/Best Configuration, delezione PPI TLK2-IRF4)*

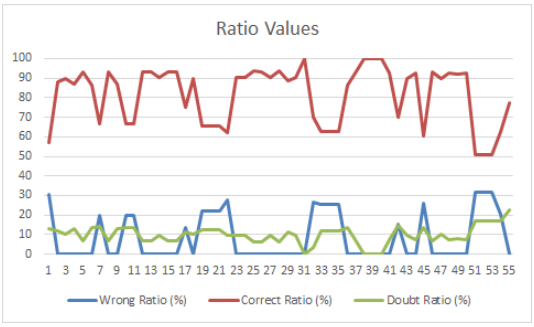
Le proteine e le interazioni proteiche validate e da segnalare risultano essere le stesse del caso precedente.

- **Validazione dei Clusters proteici**

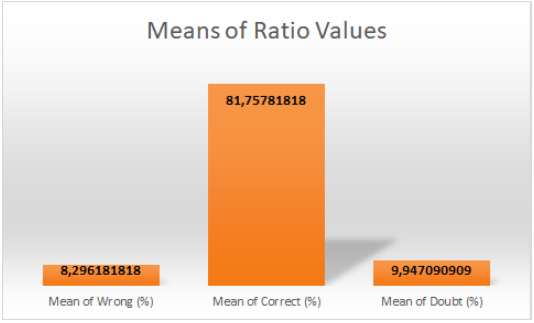
I risultati delle simulazioni per i 2 Cluster separatamente, sono mostrati in Figura 13.19 e in Figura 13.20.



(a) *Grafo processato e colorato secondo gli stati predetti*

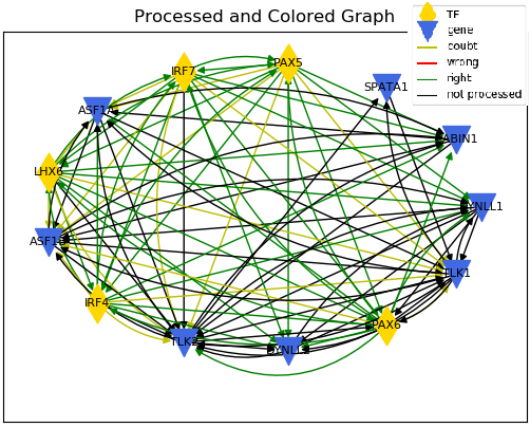


(b) *Andamento dei Valori Percentuali*

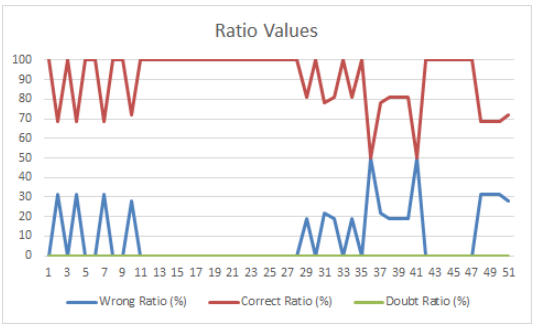


(c) *Media dei Valori Percentuali*

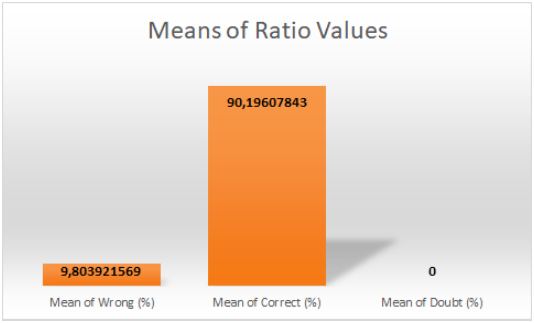
Figura 13.19: *Risultati della validazione dei Clusters proteici (Simulatore d'espressione con PPI monostato valutata, All-One/Best Configuration, delezione Bad-Performance Cluster)*



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



(c) *Media dei Valori Percentuali*

Figura 13.20: *Risultati della validazione dei Clusters proteici (Simulatore d'espressione con PPI monostato valutata, All-One/Best Configuration, delezione Good-Performance Cluster)*

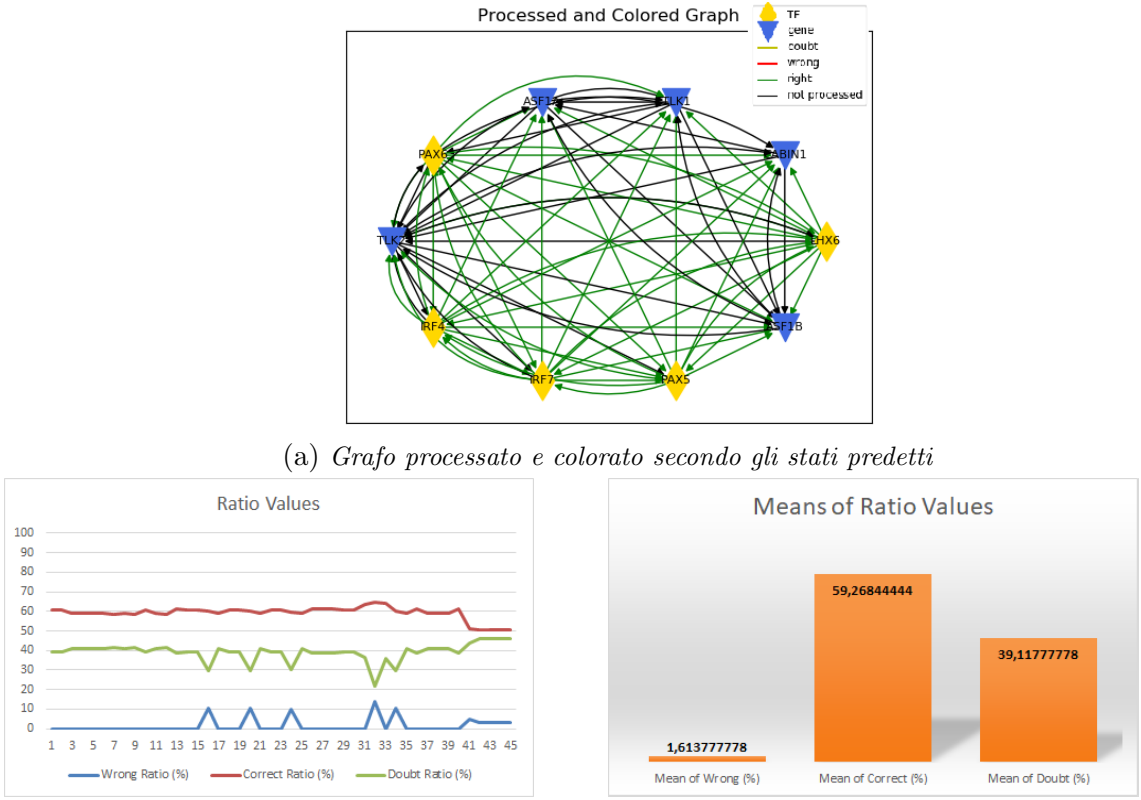
Anche queste simulazioni permettono la validazione definitiva delle proteine e delle interazioni proteiche identificate nel caso precedente.

- **Definizione dell'effetto combinato di mutazioni doppie**

Ulteriori simulazioni sono state effettuate al fine di identificare eventuali interazioni positive in mutazioni doppie appartenenti ai 2 Cluster, ossia mutazioni combinate che non apportano ulteriore danno rispetto all'eliminazione singola degli archi.
Questa condizione si può verificare quando le 2 proteine cooperano per il raggiungimento di uno scopo comune e, in tal caso, l'eliminazione di un singolo arco impedisce il raggiungimento dello scopo e, dunque, la rete non risulta maggiormente pregiudicata dall'eliminazione del secondo arco; un'altra ipotesi è che l'eliminazione della seconda interazione proteica attui una soppressione dell'effetto dovuto alla prima eliminazione.
Non sono state, però, identificate interazioni di questo tipo.

Simulatore d'espressione con PPI bistato valutata

• **Predizione del profilo d'espressione** Queste simulazioni non hanno permesso l'identificazione del Best Status dei geni interessati, in quanto i risultati ottenuti sono stati gli stessi, forzando ambedue gli stati, per tutti i geni identificati.
In Figura 13.21 sono mostrati i risultati relativi al forzamento dello stato attivo, per *ASF1A*, che risultano essere, come detto, identici a quelli relativi al forzamento dello stato inibito.



(a) *Grafo processato e colorato secondo gli stati predetti*

(b) *Andamento dei Valori Percentuali*

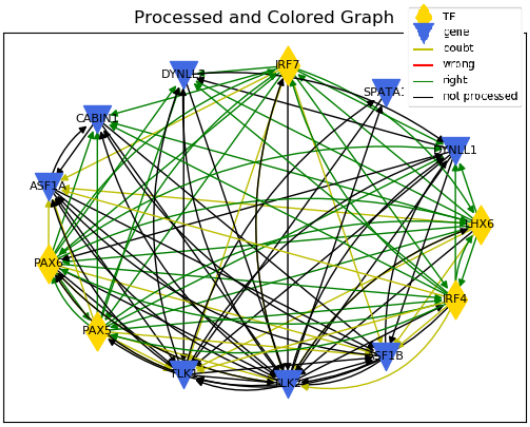
(c) *Media dei Valori Percentuali*

Figura 13.21: *Risultati della predizione del profilo d'espressione (Simulatore d'espressione con PPI bistato valutata, con ASF1A attivo)*

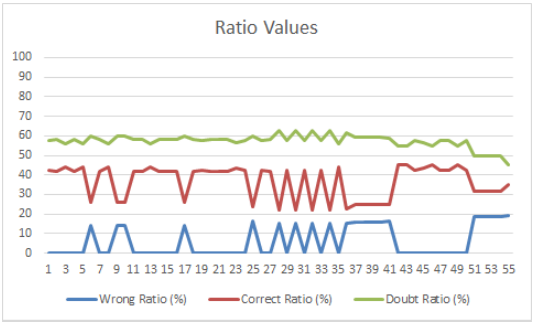
Probabilmente, questa cosa è dovuta al diretto collegamento dei geni di interesse ad archi di tipo PPI, che, in questa versione del simulatore, consentono la propagazione di ambedue gli stati possibili, impedendo una valutazione reale degli effetti della forzatura di due diversi stati, per via della mancanza di specificità.
Inoltre, come previsto, l'introduzione del doppio stato ha causato un netto incremento degli archi dubbi, che risultano essere in numero predominante rispetto alle altre due tipologie.

- **Validazione del profilo d'espressione**

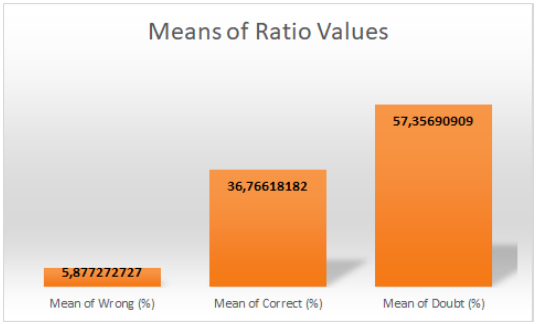
Anche in questo caso, non è stato possibile identificare la best configuration, poiché la forzatura di diversi profili di espressione, per i 3 geni di nostro interesse, ha portato agli stessi risultati.
In Figura 13.22 sono riportati i risultati relativi al profilo d'espressione di tipo *All-One*, sulla quale si basano le simulazioni successive. Questa scelta è stata fatta di modo tale da allinearsi con le informazioni presenti in letteratura.



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*

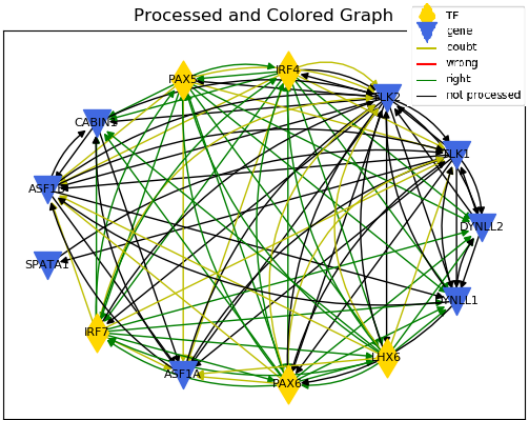


(c) *Media dei Valori Percentuali*

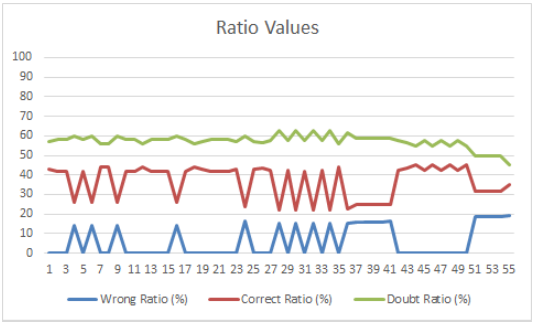
Figura 13.22: *Risultati della validazione del profilo d'espressione (Simulatore d'espressione con PPI bistato valutata, All-One Configuration)*

- **Validazione delle proteine correlate alla mutazione**

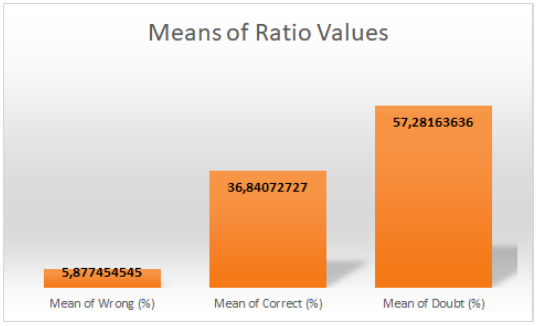
Anche in questa simulazione, sono stati identificati i due Cluster riportati precedentemente. Il Cluster di nostro interesse, ossia contenente gli archi la cui rimozione apporta un miglioramento di performance, risulta essere composto dai 4 geni identificati le altre due versioni di simulatore, ai quali, però, se ne aggiungono altri 2: *IRF4, IRF7, LHX6, PAX5 e SPATA1, TLK1*.
In Figura 13.23 sono riportate le performance dovute all'eliminazione dell'arco di PPI *TLK2-SPATA1*, che restituisce l'incremento di performance maggiori.



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



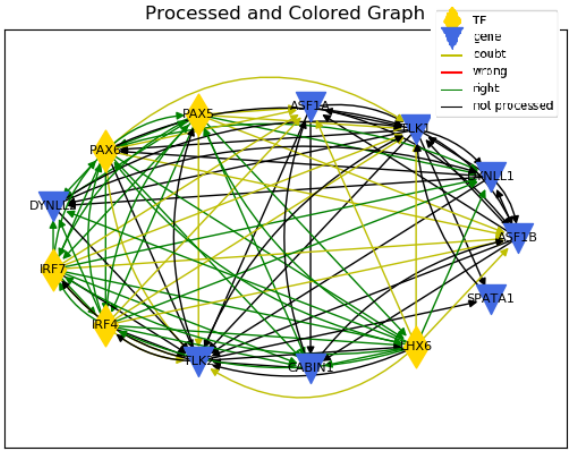
(c) *Media dei Valori Percentuali*

Figura 13.23: *Risultati della validazione delle proteine associate alla mutazione (Simulatore d'espressione con PPI bistato valutata, All-One Configuration, delezione PPI TLK2-SPATA1)*

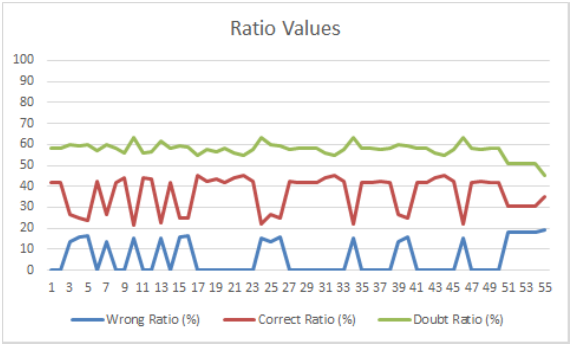
Le proteine e le interazioni proteiche da segnalare risultano essere aumentate rispetto al caso precedente.
Data la stretta correlazione tra *TLK1 e TLK2*, risulta accettabile che una mutazione coinvolgente l'interazione proteica tra queste due proteine risulti essere potenzialmente dannosa. Difficilmente comprensibile è, però, come l'eliminazione di tale interazione possa provocare una sovra-espressione dei geni in gioco, tra i quali lo stesso *TLK1*. Ciò fa pensare alla presenza di troppo rumore introdotto dall'assegnazione del doppio stato, che ha come conseguenza una diminuzione della sensibilità.
Va inoltre segnalato che la differenza di performance risulta essere minima, rispetto a quelle ottenute nei casi precedenti.
I dati si dimostrano di dubbia interpretazione e difficilmente affidabili.
In virtù di questo, le proteine e le interazioni proteiche realmente validate e segnalate, risultano essere quelle comuni ai 3 simulatori. Vengono inoltre segnalate "con riserva" anche le 2 proteine identificate mediante questo simulatore.

- **Validazione di Clusters proteici**

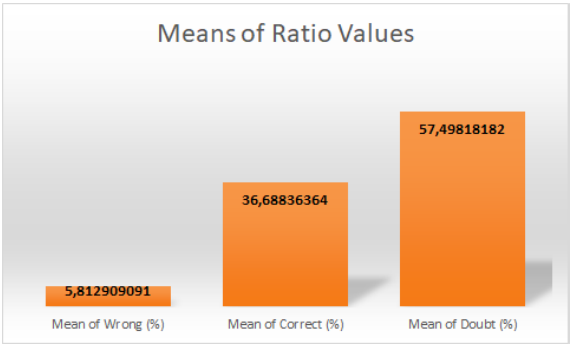
I risultati delle simulazioni per i 2 Cluster separatamente, sono mostrati in Figura 13.24 e in Figura 13.25.



(a) *Grafo processato e colorato secondo gli stati predetti*

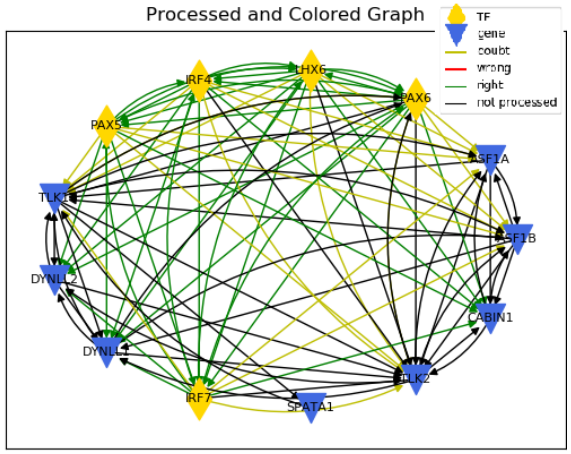


(b) *Andamento dei Valori Percentuali*

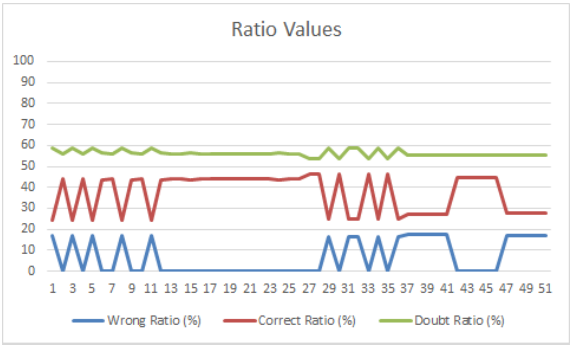


(c) *Media dei Valori Percentuali*

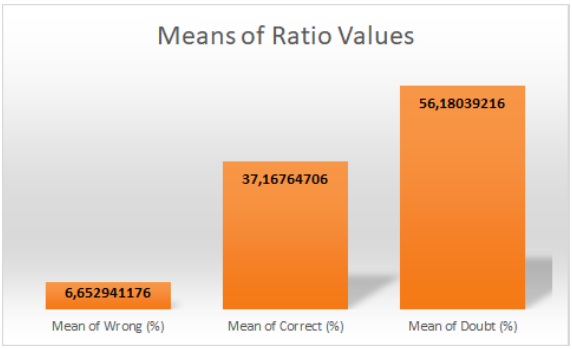
Figura 13.24: *Risultati della validazione dei clusters proteici (Simulatore d'espressione con PPI bistato valutata, All-One Configuration, delezione Bad-Performance Cluster)*



(a) *Grafo processato e colorato secondo gli stati predetti*



(b) *Andamento dei Valori Percentuali*



(c) *Media dei Valori Percentuali*

Figura 13.25: *Risultati della validazione dei clusters proteici (Simulatore d'espressione con PPI bistato valutata, All-One Configuration, delezione Good-Performance Cluster)*

Anche queste simulazioni consentono di segnalare, seppur con uno scarto di performance irrisorio le 6 proteine e le corrispondenti interazioni proteiche analizzate nella validazione precedente. Tenendo conto della scarsa specificità del simulatore introdotto e delle basse performance ottenute, le proteine segnalate in maniera definitiva risultano essere le 4 comuni a tutti i simulatori, sollevando la possibilità di includere anche le altre 2 proteine identificate con questo simulatore.

• **Definizione dell'effetto combinato di mutazioni doppie**
Anche in questo caso non sono state identificate mutazioni doppie con interazione positiva.

Parte V

Conclusioni e sviluppi futuri

Conclusioni

In questo lavoro di tesi si è partiti da una volontà nel capire come i due ambiti, la medicina e l'informatica, si sono evoluti nel tempo fino ad arrivare agli strumenti in nostro possesso.
In particolare si sono approfondite le difficoltà nelle diagnosi di persone con disturbo dello spettro autistico, fino ad arrivare a come la genetica e la biologia dei sistemi risultano essere lo strumento del futuro per l'identificazione precoce e oggettiva.
Parallelamente è stata studiata in letteratura l'evoluzione della bioinformatica, partendo dalle prime idee di genetica ai giorni nostri con la biologia dei sistemi, ambito su cui si focalizza questo lavoro.

Dalla consapevolezza del significato della biologia dei sistemi e dunque di reti di geni, si sono poi sviluppati diversi strumenti per la loro analisi.
Inizialmente si sono sviluppate cinque macro-famiglie di strumenti atti alla semplificazione e all'estrapolazione di informazioni: una basata sulla divisione di proprietà presenti all'interno degli elementi contenuti in TheRingDB, uno database innovativo che oltre a riunire più database introduce una standardizzazione all'interno; una basata sullo studio a livello proteico, con un'estrapolazione e un'analisi dei complessi; una che esplora i diversi cammini tra gli elementi; una che sfrutta un'analisi basata sulla componente fortemente connessa e infine una che si basa sulla variabile informatica di betweennes.
Tutti questi strumenti sono stati sviluppati avendo sempre in mente il significato biologico del risultato.

In seguito, sono stati costruiti altre tecniche di analisi di reti di geni, che prevedono la propagazione delle espressioni nelle reti di regolazione e non.
In particolare sono stati definiti due tipi di simulatore: uno basato su un confronto con un Golden Standard e l'altro con forzamento dell'espressione.
Nel primo caso, il simulatore propaga l'attivazione o l'inibizione della sorgente fino all'elemento target e poi confronta il risultato della propagazione di ogni singolo nodo con un file contenente tutte le espressioni degli elementi. Questo simulatore è stato progettato anche con una possibile interazione dell'utente, il quale può decidere di eliminare relazioni tra elementi e verificarne l'effetto causato.
Nel secondo invece, è stato pensato di forzare l'espressione di un pool di geni e di quantificare il grado di coerenza della rete in relazione a ciò. Il grado di coerenza viene rappresentato attraverso percentuali di giusti, sbagliati e dubbi, per ogni arco analizzato all'interno della rete. Inoltre, è stato assegnato anche un valore predetto a ogni arco sfruttando delle soglie. Il valore predetto corrisponde a due stati: giusto o errato, la cui definizione dipende dalle percentuali di errati, dubbi o corretti.
Di questo ultimo simulatore sono state fatte quattro versioni, una che tiene conto solo della rete regolatoria; una che tiene conto anche della rete a livello di proteoma ma conferendo all'elemento proteico il valore d'espressione del gene appartenente alla rete di regolazione; un'altra che oltre a mantenere le interazioni proteiche valuta il valore di dubbio sull'espressione della proteina senza conferirgli lo stato del complesso; infine una che considera le tre possibilità del complesso proteico, cioè la totale attivazione, la totale inibizione e la parziale attivazione/inibizione.

Successivamente, i ricercatori dell'Università di Torino hanno fatto recapitare una lista di geni coinvolti nell'autismo, più specificatamente un elenco di 126 elementi che interagiscono con il gene TLK2 mutato a livello di interattoma. Lo studio di tutti gli elementi combinati si è rivelato però molto oberante per quanto riguarda la potenza ma anche i tempi computazionali, pertanto sono state fatte delle restrizioni delle relazioni andando a non considerare i cammini al di sopra di un fattore due, ciò con più di tre oggetti, operazione che ha portato alla perdita di tre elementi.
Ogni cammino tra i singoli 123 elementi e il gene TLK2 sono stati analizzati, per poi decidere di focalizzarci su 3 geni con un interesse maggiore dal punto di vista dello spettro autistico: TLK1, ASF1A e ASF1B.
Il gene TLK2 è strettamente correlato al disturbo dello spettro autistico, lo dimostrano numero studi al riguardo e la sua presenza nel Autism Database; nello stesso tempo anche gli altri 3 geni, oltre ad essere strettamente relazionati al gene TLK2, sono in via sperimentale correlati allo stesso disturbo.
Concentrandoci su questo pool ridotto di quattro elementi sono stati prodotti i risultati sul secondo simulatore sviluppato, in quanto è quello che permette l'analisi dei dati in nostro possesso.
I risultati sono stati ottenuti, con tutti e quattro i tipi di versioni del simulatore, sia studiando solo i cammini che congiungono TLK2 a uno degli altri tre elementi, sia analizzando la rete contenente i quattro e gli altri elementi di congiungimento. Questi risultati hanno predetto e validato, in due versioni del simulatore, il profilo d'espressione dei nostri geni d'interesse, che ne prevede l'attivazione.
Successivamente sono state fatte delle simulazioni di una mutazione del gene TLK2, andando a eliminare archi uscenti da esso, riproducendo così una perdita di funzionalità. L'eliminazione delle relazioni è stata fatta sia di un singolo arco per volta, andando ad analizzare l'effetto di una sola mutazione, sia di più archi contemporaneamente, riproducendo l'effetto di mutazioni clusterizzate e analizzando l'effetto della combinazione doppia. Ciò ha permesso di validare, in tutti i simulatori, le proteine codificate dai quattro geni collegati direttamente a TLK2 (PAX5, IRF4, IRF7 e LHX6), e le corrispettive interazioni proteiche con lo stesso. L'ultimo simulatore ha permesso di segnalare, seppur con riserva, altre due proteine codificate dai corrispettivi geni (SPATA1 e TLK1).
Le proteine identificate saranno segnalate ai ricercatori dell'Università di Torino, che provvederanno a indagarle, con la speranza di trovare un nesso con lo sviluppo dell'autismo.

Sviluppi futuri

Nel corso di questo progetto sono stati sviluppati numerosi tool di diverso genere per poter fornire degli strumenti utili nello studio di reti biologiche.
Nella Sezione II sono stati illustrate diverse metodologie per l'estrapolazione di informazioni chiare e puntuali, alcune sulla base della semplice analisi dello standard utilizzato, altre su variabili informatiche con un nesso biologico; ma tutte queste strategie derivano da uno studio di sole due persone e durante un tempo limitato, quindi è molto probabile che in un futuro si possa espandere questa sezione con più funzionalità, anche più elaborate. Le variabili informatiche sono numerosissime e l'estrazione degli oggetti diventa di un numero spropositato se si pensa a tutte le combinazioni possibili, per cui con certezza si avrà un'espansione, ma si spera sempre basandosi su una spiegazione biologica dei risultati.

Nella Sezione III invece, sono stati sviluppati dei simulatori di reti.
In particolare nel Capitolo 10 è stato costruito un prototipo basato sul confronto con un Golden Standard.Il prototipo potrebbe avere un futuro se solo si avessero a disposizione i dati, cioè le espressioni di tutti i nodi o entità biologiche, per cui lo sviluppo in questo caso è rivolto principalmente alla ricerca biologica. La difficoltà nell'ottenimento di questi dati non è per nulla da sottovalutare, in quanto le espressioni tendono a cambiare non solo da soggetto a soggetto, ma anche all'interno della giornata dello stesso, in quanto sono molto sensibili ai ritmi circadiani. Una possibile strategia sarebbe avere una media, quindi non un numero binario come 0 o 1, e in questo caso il simulatore dovrebbe subire una leggera modifica nei filtri, ponendo ad esempio un valore maggiore di una certa soglia piuttosto che eguagliarlo a uno dei due stati binari. Tutto ciò rende il futuro di questo simulatore molto difficile, ma in un avvenire sarà sicuramente possibile.
Avendo a disposizione l'informazione sulle espressioni, il confronto con la propagazione del simulatore sarebbe possibile e inoltre aprirebbe una strada di miglioramenti dello stesso. Un miglioramento che potrebbe essere applicato è senza alcun dubbio l'automatizzazione nella rimozione degli archi per vedere gli effetti, azione che per il momento viene svolta manualmente dall'utente con la scelta degli archi da rimuovere.

Nel Capitolo 11 è stato presentato un prototipo di simulatore che, forzando lo stato d'espressione di un pool di geni, permette di ottenere una valutazione della coerenza di ogni arco. Di questo prototipo sono state fornite 4 versioni definitive, che permettono di elidere o processare archi di natura diversa dalla regolatoria, ossia di interazione proteica, in maniera diversa. Ogni arco PPI, però, viene trattato in modo superficiale e spesso impreciso (si forza una correlazione tra la presenza/assenza della proteina e lo stato espressivo del gene a monte), per via della mancanza di effettivi dati di espressione, che potrebbero arricchire l'informazione relativa al rapporto tra le entità.
Inoltre, la mancanza dei dati relativi al profilo d'espressione dei geni di interesse, ha richiesto delle strategie di predizione e di validazione del probabile stato degli stessi.

Ulteriore limite è rappresentato dal costo computazionale delle simulazioni, che ha richiesto l'estrazione di un network limitato ai soli percorsi, presenti tra i vari geni, di lunghezza pari a 2, nonché un nuovo taglio della rete, necessario in fase di simulazione. L'integrazione in un Server, caratterizzato da una potenza di calcolo maggiore, potrebbe permettere l'estrazione di un network di partenza maggiore e, consecutivamente, maggiormente accurato ed eliminare ogni necessità di effettuare tagli e riduzioni dello stesso, durante le simulazioni.

Bibliografia

[1] Jon Baio Lisa Wiggins Deborah L. Christensen et al. «Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014». In: *MMWR- Morbidity and Mortality Weekly Report* 67 No.6 (27 apr. 2018). DOI: 10.15585/mmwr.ss6706a1. URL: https://www.ncbi.nlm.nih.gov/pubmed/29701730.

[2] Pavlopoulos Secrier Moschopoulos et al. «Using graph theory to analyze biological networks». In: *BioData Min.* 4(1) (2011). DOI: 10.1186/1756-0381-4-10.

[3] West D. B. *Introduction to Graph Theory - Second Edition*. A cura di Pearson Education Inc. 2001. ISBN: 81-7808-830-4.

[4] Adriana Birlutiu. «Machine learning for pairwise data : applications for preference learning and supervised network inference». In: *International Economics and Economic Policy* (gen. 2011).

[5] Anastasis Oulas George Minadakis Margarita Zachariou Kleitos Sokratous Marilena M. Bourdakou e George M. Spyrou. «Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches». In: *Briefings in Bioinformatics* (27 nov. 2017), pp. 806–824. DOI: 10.1093/bib/bbx151. URL: https://academic.oup.com/bib/article-abstract/20/3/806/4662946.

[6] Marc Colet e Robert Herzog. «WWW2GCG, a web interface to the GCG biological sequences analysis software». In: *Computers and Graphics* 20(3) (1 mag. 1996), pp. 445–450. DOI: 10.1016/0097-8493(96)00014-3. URL: https://www.sciencedirect.com/science/article/abs/pii/0097849396000143.

[7] Xosé M Fernández Daniel J Rigden. «The 2018 Nucleic Acids Research database issue and the online molecular biology database collection». In: *Nucleic Acids Research* 46(D1) (4 gen. 2018), pp. D1–D7. DOI: 10.1093/nar/gkx1235. URL: https://academic.oup.com/nar/article/46/D1/D1/4781210.

[8] Chang Dayhoff Eck e Sochard. «ATLAS of PROTEIN SEQUENCE and STRUCTURE 1965». In: (1965). A cura di National Biomedical Research Foundation.

[9] Center for Disease Control e prevention. «Data and Statistics on Autism Spectrum Disorder». In: (3 set. 2019). A cura di U.S. DEPARTMENT OF HEALTH e HUMAN SERVICES. URL: https://www.cdc.gov/ncbddd/autism/data.html.

[10] The Simon Foundation. *SFARI Database*. URL: https://gene.sfari.org/database/ring-browser/.

[11] Chartrand G. e Lesniak L. *Graphs and Digraphs – Second Edition*. A cura di California Wadsworth In Belmont. 1989. ISBN: 0-543-06324-1.

[12] S. Di Carlo G. Politano e A. Benso. «One DB to rule them all'—the RING: a Regulatory INteraction Graph combining TFs, genes/proteins, SNPs, diseases and drugs». In: *Database (2018)* 2019.Article ID baz108 (4 nov. 2019), pp. 1–14. ISSN: 1758-0463. DOI: 10.1093/database/baz108. URL: http://dx.doi.org/10.1093/database/baz108.

[13] Mrs. Kanchan Gawande e Mr. Dhiraj Rane. «Exploring the Applications and Potential of Bioinformatics». In: *IOSR Journal of Computer Engineering (IOSR-JCE)* (2016), pp. 20–26.

[14] Swart P. Hagberg A. Schult D. «Networkx Reference – Release 2.2». In: (2018).

[15] National Institutes of Health. *Autism Spectrum Disorder*. A cura di U.S. DEPARTMENT OF HEALTH e HUMAN SERVICES. 2018. DOI: 19-MH-8084. URL: https://www.nimh.nih.gov/health/publications/autism-spectrum-disorder/19-mh-8084-autismspectrumdisorder_152236.pdf.

[16] A. D. Hershey e Martha Chase. *Independent functions of viral protein and nucleic acid in growth of bacteriophage*. 20 Set. 1952. DOI: 10.1085/jgp.36.1.39. URL: http://jgp.rupress.org/content/jgp/36/1/39.full.pdf.

[17] Silljé HH e Nigg EA. «Identification of human Asf1 chromatin assembly factors as substrates of Tousled-like kinases». In: *Current Biology* 11(13) (10 lug. 2001), pp. 1068–1073. DOI: 10.1016/S0960-9822(01)00298-6. URL: https://www.cell.com/current-biology/fulltext/S0960-9822(01)00298-6?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0960982201002986%3Fshowall%3Dtrue.

[18] H H Silljé K Takahashi K Tanaka G Van Houwe e E A Nigg. «Mammalian homologues of the plant Tousled gene code for cell-cycle-regulated kinases with maximal activities linked to ongoing DNA replication». In: *The EMBO journal* 18(20) (15 ott. 1999), pp. 5691–5702. DOI: 10.1093/emboj/18.20.5691. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1171636/.

[19] Tanecious Hunter. *The Code of Life*. A cura di Future Crunch. 3 Ago. 2017. URL: https://futurecrun.ch/articles/the-code-of-life.

[20] Steve J. Charette Jeff Gauthier Antony T. Vincent e Nicolas Derome. «A brief history of bioinformatics». In: *Briefings in Bioinformatics* 5(2) (2018), pp. 1–16. DOI: 10.1093/bib/bby063. URL: https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bby063/5066445.

[21] Sal Khan. *Introduction to heredity: Mendel and his peas*. A cura di Khan Academy. 30 Set. 2009. URL: https://www.khanacademy.org/science/high-school-biology/hs-classical-genetics/hs-introduction-to-heredity/a/mendel-and-his-peas.

[22] Klimovskaia e Young et al. «Tousled-like kinases phosphorylate Asf1 to promote histone supply during DNA replication». In: *Nature Communications* 5 (6 mar. 2014). DOI: 10.1038/ncomms4394. URL: https://www.nature.com/articles/ncomms4394#citeas.

[23] Avi Ma'ayan. «Introduction to Network Analysis in Systems Biology». In: *Science Signaling* (2011). DOI: 10.1126/scisignal.2001965.

[24] M Younus Wani NA Ganie S Rani S Mehraj MR Mir MF Baqual KA Sahaf FA Malik e KA Dar. «Advances and applications of Bioinformatics in various fields of life». In: *International Journal of Fauna and Biological Studies* 5(2) (5 feb. 2018), pp. 03–10. URL: http://www.faunajournal.com/archives/2018/vol5issue2/PartA/5-1-52-692.pdf.

[25] PyData Development Team McKinney W. «pandas: powerful Python data analysis toolkit – Release 0.24.2». In: (2019).

[26] Leila McNeill. *How Margaret Dayhoff Brought Modern Computing to Biology*. A cura di SMITHSONIAN.COM. 9 Apr. 2019. URL: https://www.smithsonianmag.com/science-nature/how-margaret-dayhoff-helped-bring-computing-scientific-research-180971904/.

[27] NA. *ASF1A*. A cura di Wikipedia-The free encyclopedia. 29 Ago. 2017. URL: https://en.wikipedia.org/wiki/ASF1A.

[28] NA. *ASF1B*. A cura di Wikipedia-The free encyclopedia. 29 Ago. 2017. URL: https://en.wikipedia.org/wiki/ASF1B.

[29] NA. *Bioinformatica*. A cura di l'enciclopedia libera Wikipedia. 9 Lug. 2019. URL: https://it.wikipedia.org/wiki/Bioinformatica.

[30] NA. *Bioinformatics (journal)*. A cura di l'enciclopedia libera Wikipedia. 21 Lug. 2019. URL: https://en.wikipedia.org/wiki/Bioinformatics_(journal).

[31] NA. *Frederick Sanger*. A cura di Wikipedia-The free encyclopedia. 31 Ott. 2019. URL: https://en.wikipedia.org/wiki/Frederick_Sanger.

[32] NA. *NAR Database Summary Paper Category List*. A cura di Oxford Academic Journals. URL: http://www.oxfordjournals.org/nar/database/c/.

[33] NA. *Nucleosoma*. A cura di Wikipedia-The free encyclopedia. 27 Nov. 2018. URL: https://it.wikipedia.org/wiki/Nucleosoma.

[34] NA. *TLK1*. A cura di Wikipedia-The free encyclopedia. 4 Set. 2019. URL: https://en.wikipedia.org/wiki/TLK1.

[35] NA. *TLK2*. A cura di Wikipedia-The free encyclopedia. 4 Lug. 2019. URL: https://en.wikipedia.org/wiki/TLK2.

[36] NA. *Types of biological networks*. A cura di Train online. URL: https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/networks-cell-biology-summary-0.

[37] omicsouts, cur. *Disease and systems biology*. URL: https://omicscouts.com/en/disease-and-systems-biology.html.

[38] Peipei e Yuan et al. «BioID: A Proximity-Dependent Labeling Approach in Proteomics Study». In: *Methods in Molecular Biology* (2019). DOI: 10.1007/978-1-4939-8814-3_10. URL: https://www.ncbi.nlm.nih.gov/pubmed/30276738.

[39] Corrado Priami. «Informatica e biologia dei sistemi». In: *Mondo digitale n.1* (mar. 2004). URL: http://domino.aicanet.it/dev/reposit.nsf/.

[40] Reijinders e Miller et al. «De Novo and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype-Phenotype Evaluation of a Distinct Neurodevelopmental Disorder». In: *The American Journal of Human Genetics* 102 (7 giu. 2018), pp. 1195–1203. DOI: 10.1016/j.ajhg.2018.04.014. URL: https://www.sciencedirect.com/science/article/pii/S0002929718301617.

[41] Roux e Dae et al. «BioID: A Screen for Protein-Protein Interactions». In: *Curr Protoc Protein Sci. Author manuscript; available in PMC* (21 feb. 2019). DOI: 10.1002/cpps.51. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6028010/.

[42] Silberman S. *NeuroTribù – I talenti dell'autismo e il futuro della neurodiversità*. A cura di Italia Edizioni LSWR Milano. 2016. ISBN: 978-88-6895-360-7.

[43] Segura-Bayona e Stracker. «The Tousled-like kinases regulate genome and epigenome stability: implications in development and disease». In: *Cellular and Molecular Life Sciences* (24 giu. 2019). DOI: 10.1007/s00018-019-03208-z.

[44] Takahashi et al. Silljé. «Mammalian homologues of the plant Tousled gene code for cell-cycle-regulated kinases with maximal activities linked to ongoing DNA replication». In: *The EMBO journal* 18 (20 1999), pp. 5691–5702.

[45] Jochen Singer Anja Irmisch Hans-Joachim Ruscheweyh Franziska Singer Nora C. Toussaint Mitchell P. Levesque Daniel J. Stekhoven e Niko Beerenwinkel. «Bioinformatics for precision oncology». In: *Briefings in Bioinformatics* 20(3) (18 dic. 2018), pp. 778–788. DOI: 10.1093/bib/bbx143. URL: https://academic.oup.com/bib/article/20/3/778/4758621.

[46] Grandin T. e Panck R. *Il Cervello Autistico*. A cura di Italia Adelphi Edizioni Milano. 2014. ISBN: 978-88-459-2894-9.

[47] N-E. Eriksson T.K. Attwood A. Gisel e E. Bongcam-Rudloff. «Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective». In: *Bioinformatics - Trends and Methodologies, Dr. Mahmood A. Mahdavi (Ed.)* (2 Nov. 2011). DOI: 10.5772/786. URL: https://www.intechopen.com/books/bioinformatics-trends-and-methodologies/concepts-historical-milestones-and-the-central-place-of-bioinformatics-in-modern-biology-a-european.

[48] Fergus Walsh. *The most important photo ever taken?* A cura di BBC News. 16 Mag. 2012. URL: https://www.bbc.com/news/health-18041884.

[49] J. Thompson G. Braun D. Tierney L. Wessels e H. Schmitzer. «Rosalind Franklin's X-ray photo of DNA as an undergraduate optical diffraction experiment». In: *American Journal of Physics* 86(2) (18 gen. 2017). DOI: 10.1119/1.5020051. URL: https://aapt.scitation.org/doi/10.1119/1.5020051.

Ringraziamenti

Ci terremo a ringraziare innanzitutto i Professori Alfredo Benso e Gianfranco Michele Maria Politano, che ci hanno seguite dandoci l'opportunità di svolgere questo lavoro di tesi.
Ulteriore menzione va alla Professoressa Gabriella Olmo e al Professor Stefano Di Carlo.
Vorremmo ringraziare infine tutti i professori che hanno contribuito alla nostra formazione, rendendoci parte di quello che siamo.
Ringraziamo le nostre famiglie e tutte le persone che ci hanno sostenuto durante questo percorso.