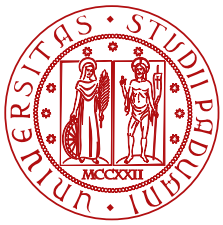

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

Analisi delle Traiettorie Conservate a Livello di Pathway in Alberi Filogenetici Tumorali

Relatore
Prof. Vandin Fabio

Laureando
Bastianello Mattia

Correlatore
Pellegrina Leonardo

ANNO ACCADEMICO 2023-2024

Data di laurea 27/09/2024

“Io sono quello che sono, e non me ne vergogno. “Non ti vergognare mai”, disse mio padre,
“c’è qualcuno che ce l’avrà con te, ma non ne vale la pena”
(Rubeus Hagrid - Harry Potter e il Calice di Fuoco)

Sommario

I tumori derivano da un processo evolutivo che porta alla formazione di colonie cellulari, chiamate cloni, caratterizzate da significative variazioni genetiche. Comprendere le traiettorie evolutive conservate all'interno degli alberi filogenetici di diverse colonie tumorali è cruciale per approfondire la nostra conoscenza delle mutazioni che portano alla genesi dei tumori. Questo approccio consente una rilevazione e previsione più rapida dei progressi nella progressione tumorale. MASTRO è un algoritmo che analizza un insieme di alberi filogenetici relativi a una coorte di tumori ed esegue un'analisi statistica delle traiettorie identificate, permettendo di selezionare quelle più significative e scartare le meno rilevanti. Questa tesi affronta il problema di analizzare le traiettorie tumorali a livello di pathway partendo della conversione dei dati in ingresso per l'algoritmo MASTRO, trasformando gli alberi genetici in alberi di pathways attraverso diversi criteri di importanza. Viene descritto il processo di sviluppo di un programma progettato per eseguire queste conversioni in modo preciso. La tesi include un'analisi approfondita e un confronto dei risultati ottenuti tramite MASTRO, evidenziando le differenze tra i risultati ottenuti dai dataset convertiti e da quelli originali.

Indice

Elenco delle figure

Capitolo 1

Introduzione

1.1 Definizione del problema

Il processo di formazione dei tumori è caratterizzato dall'accumulo di alterazioni somatiche che conferiscono vantaggi selettivi a una o più cellule. Questo accumulo di mutazioni porta a un'eterogeneità genotipica all'interno delle cellule tumorali. Tuttavia, esistono alcune caratteristiche condivise, come l'ordine di insorgenza di specifiche mutazioni, che sono comuni nella maggior parte delle cellule di un determinato gruppo tumorale. Identificare queste caratteristiche comuni è di grande importanza, poiché può facilitare una diagnosi precoce dei tumori e consentire una previsione più accurata della loro progressione.

Le alterazioni geniche sono associate a vari pathways, e spesso più geni sono mappati sugli stessi pathways. La possibilità di convertire queste rappresentazioni in alberi di pathways consente di analizzare e interpretare queste mutazioni da una prospettiva diversa, offrendo una visione più integrata dei processi biologici coinvolti.

1.2 MASTRO e formato dei dati in input e in output

MAximal tumor treeS TRajectOries (MASTRO)[2] è un algoritmo per scoprire traiettorie evolutive significativamente conservate nei tumori. MASTRO identifica tutte le traiettorie conservate in una raccolta di alberi filogenetici che descrivono l'evoluzione di una coorte di tumori, permettendo di individuare relazioni complesse e conservate tra le alterazioni. MASTRO valuta la significatività delle traiettorie utilizzando un test statistico condizionale che cattura la coerenza nell'ordine in cui le alterazioni vengono osservate in diversi tumori, si rimanda alla Figura 1.1 per una descrizione a livello superiore del funzionamento di MASTRO.
MASTRO richiede in input un insieme di alberi filogenetici che descrivono le evoluzioni di tumori di un gruppo di cellule, poi produce come output tutte le traiettorie osservate in alme-

no σ tumori, dove questa soglia è scelta dall'utente al momento dell'esecuzione dell'algoritmo. E' importante sottolineare che questi alberi descrivono tutte le mutazioni subite da una cellula, non è assunto che siano consecutive, ma solo che appaiano nell'ordine indicato dallo specifico albero.

1.2.1 Concetti chiave e definizioni fondamentali per MASTRO

L'input per l'algoritmo MASTRO è un multinsieme di n alberi tumorali radicati, ottenibili mediante diversi metodi computazionali che ricostruiscono la storia evolutiva dei tumori corrispondenti utilizzando dati bulk o dati a singola cellula. Ogni nodo dell'albero rappresenta un clone tumorale ed è associato ad una collezione di alterazioni geniche.
La radice di ogni albero contiene l'insieme vuoto e rappresenta le cellule normali (germinali), mentre ogni nodo non radice contiene un sottoinsieme non vuoto di alterazioni provenienti da un insieme complessivo di m alterazioni. Le alterazioni presenti in un nodo sono quelle che appaiono nel clone corrispondente ma non nei suoi antenati, mentre l'insieme totale delle alterazioni di un clone è dato dall'unione degli insiemi di alterazioni trovati nel percorso unico che va dal nodo corrispondente alla radice. Si assume che ogni alterazione appaia al massimo una volta in ciascun albero, anche se l'input può comunque rappresentare eventi distinti sulla stessa regione genomica.
Per qualsiasi albero tumorale, si definisce una alterazione a, contenuta nel nodo v, è un'antenata dell'alterazione b, contenuta nel nodo w, se il nodo v appartiene al percorso che va da w alla radice dell'albero tumorale.
Definiamo una traiettoria come un albero tumorale T. Una traiettoria τ è osservata in un albero T' se l'insieme delle alterazioni contenute in τ è un sottoinsieme delle alterazioni presenti in T' e se tutti gli ordinamenti temporali tra coppie di alterazioni in τ sono rispettati in T'. Formalmente, diciamo che la traiettoria τ è osservata in un albero T' se le seguenti condizioni sono soddisfatte:

1. per ogni coppia di alterazioni a e b in τ tale che a è un antenato di b in τ, allora a è un antenato di b in T';

2. per tutte le coppie di alterazioni a e b di τ che appartengono allo stesso nodo in T', queste rimangono nello stesso nodo in T' (l'ordinamento tra loro non è noto);

3. per tutte le coppie di alterazioni a e b di τ che si trovano in rami diversi di T', esse appartengono a rami differenti in T'.

Un'altra rappresentazione equivalente è data dal grafo tumorale espanso G di un albero tumorale T. Questo grafo diretto è costruito come segue:

1. per ogni alterazione *a* in *T* contenuta in un nodo *v*, esiste un nodo v_a in *G* che contiene solo l'alterazione *a*;

2. per ogni coppia di alterazioni *a* e *b* in *T*, esiste un arco diretto da v_a a v_b se e solo se *a* è un antenato di *b* in *T*;

3. per ogni coppia di alterazioni *a* e *b* appartenenti allo stesso nodo in *T*, esiste un arco speciale non diretto v_a,v_b in *G*, dove l'etichetta dell'arco indica che l'ordinamento tra *a* e *b* è sconosciuto

4. *G* contiene un nodo vuoto v_r (la radice di *T*) e un arco diretto da v_r a tutti gli altri nodi di *G*

Nella figura 1.2 vengono mostrati esempi di alberi tumorali e dei loro grafi tumorali espansi.

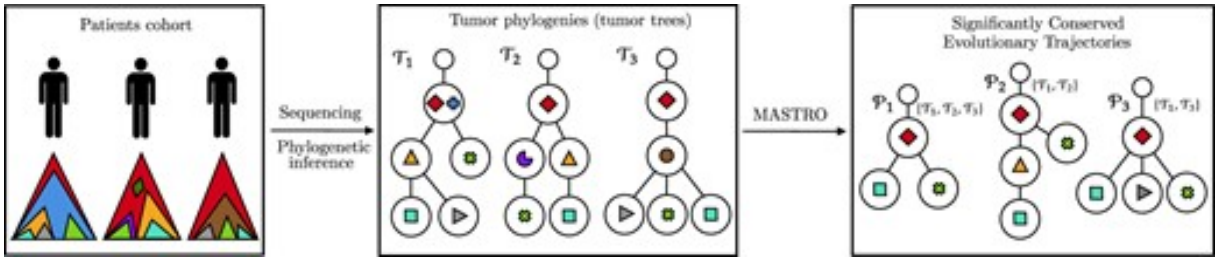

Figura 1.1: Descrizione ad alto livello del funzionamento di MASTRO.
Figura presa da [2].

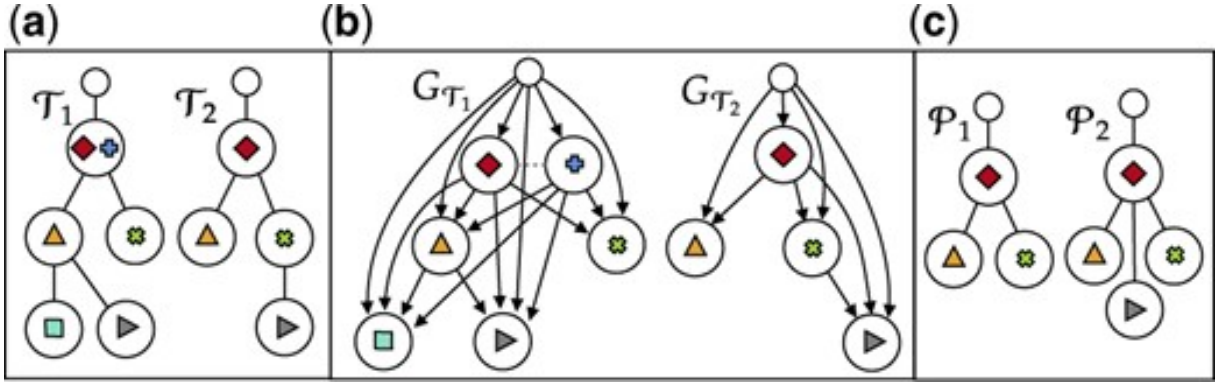

Figura 1.2: **(a)** Alberi tumorali T_1 e T_2. **(b)** Grafi tumorali espansi G_{T_1} e G_{T_2} (presente anche l'arco non diretto rappresentato dai trattini).**(c)** Traiettorie trovate come sotto grafi di G_{T_1} e G_{T_2}.
Figura presa da [2].

1.2.2 Formato di input dei file

Nella pratica l'elaborazione di MASTRO si basa su un documento di testo in cui ogni riga rappresenta l'insieme degli archi del grafo espanso associato ad una filogenia tumorale. Ogni riga del file contiene una serie di coppie di nodi collegate attraverso tre tipologie di archi:

- **A->-B** indica che il nodo A è un **antenato** di B
- **A-/-B** indica che i nodi A e B appartengono a **rami distinti**, pertanto nessuno dei due è antenato dell'altro
- **A-?-B** indica che l'ordine tra i nodi A e B non è noto

È fondamentale sottolineare che, all'interno di ogni albero, non possono esistere cicli del tipo:

```
A->-B B->-A
```

Inoltre, poiché ogni arco può apparire al massimo una volta nel grafo, il seguente esempio non è corretto:

```
A->-B A->-B
```

Di seguito è presentato un esempio di come può essere rappresentato, in un file adatto per l'elaborazione di MASTRO, il seguente grafo tumorale espanso:

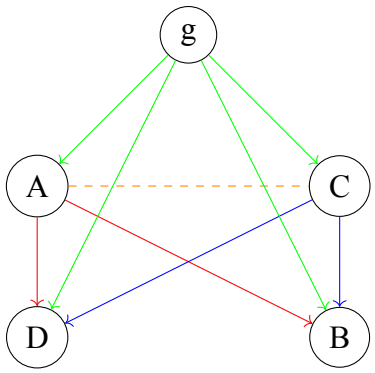

Figura 1.3: Esempio di grafo tumorale espanso

Di seguito è riportata la riga che rappresenta il grafo mostrato nella Figura 1.3:

```
A-?-C A->-B A->-D C->-B C->-D B-/-D
```

1.3 Dataset utilizzati

L'analisi svolta in questa tesi tramite MASTRO utilizzerà due dataset: uno riguarda i dati sulla leucemia mieloide acuta (AML), in cui sono presenti 120 alberi filogenetici, mentre l'altro

riguarda i dati relativi al carcinoma polmonare non a piccole cellule (NSCLC) che contiene 89 alberi filogenetici. La differenza tra i due dataset è data oltre al numero di alberi, anche dalla differenza di distribuzione dei nodi negli alberi dei dataset. Nei dati dell'AML, la maggior parte degli alberi presenta più di 3 nodi, con una media di 3,42 nodi per albero. Al contrario, negli alberi del NSCLC, il numero massimo di nodi (compreso il nodo germinale) è di 5, con una media di 2,7 nodi per albero. Questa differenza è dovuta al fatto che, nel NSCLC, la maggior parte delle alterazioni è osservata con una frequenza molto alta e, pertanto, non è ordinata in modo affidabile.

A questi due dataset si aggiungeranno ulteriori dataset ottenuti tramite le due conversioni che verranno descritte successivamente. In questi nuovi dataset, gli alberi filogenetici saranno sostituiti da alberi dei pathway, generati attraverso una mappatura gene-pathway, sempre mantenendo il formato di input richiesto dall'algoritmo MASTRO per l'analisi dei dati.

Capitolo 2

Algoritmo di conversione

In questo capitolo, verrà analizzato l'algoritmo sviluppato per questa tesi, il quale consente di convertire dataset contenenti alberi genici, espressi nel formato richiesto in input da MASTRO, in dataset che, attraverso una mappatura che assegna a ciascun gene uno o più pathways, generano così nuovi alberi. Le caratteristiche di questi alberi saranno descritte in seguito.

Le istruzioni dettagliate per l'esecuzione del programma sono disponibili nel file README della pagina GitHub [1]. Questa pagina contiene anche il codice sorgente completo, insieme ai risultati delle conversioni e delle analisi eseguite.

2.1 Struttura del codice

Il codice è composto principalmente da *run_converter.py*, da *close_to_root_converter.py* e *rnd_int_node_converter.py*. Il primo programma prende in input il file da convertire e lo passa come parametro ai due modelli di conversione. In output verrà ottenuta la conversione degli alberi secondo i due metodi di punteggio e di importanza dei geni.

Di seguito è presentato lo pseudocodice dell'algoritmo generale di conversione:

Algorithm 1 Conversione Gene-to-Pathway

```
for tree = 1, 2, ..., N do
    for edge = 1, 2, ..., M do
        update_score(node_dict, edge)        ▷ choose from different types of scoring policy
    end for
    pathway_importance_criterio(node_dict)
    Create converter with significant genes
    for edge = 1, 2, ..., M do
        create_edge(converter, first_node, second_node, edge, temp)
    end for
end for
```

Nel contesto della presente analisi, per ciascun albero contenuto nel file, viene effettuata un'operazione di aggiornamento del punteggio di ogni gene secondo il criterio prestabilito. Completata questa fase, si procede alla creazione di un dizionario in cui vengono selezionati e filtrati i pathways associati ai geni di maggiore rilevanza. Alla fine del processo, viene generato un nuovo albero basato sulle informazioni aggiornate. Le funzioni impiegate per tali operazioni sono dettagliate successivamente.

La funzione *update_score(node_dict, edge)* e *pathway_importance_criterio(node_dict)* rappresentano l'aggiornamento e la creazione del dizionario filtrato e verranno esaminate nei paragrafi successivi.

La funzione *create_edge(converter, first_node, second_node, edge, temp)* accetta come parametri di ingresso un dizionario filtrato che associa i pathways ai geni di maggiore importanza, l'arco da convertire (composto dal primo nodo, dal secondo nodo e dal tipo di arco rilevato), e l'albero in fase di costruzione. La funzione si occupa di convertire i due nodi secondo i pathways associati: nel caso in cui un nodo sia associato a più pathways, questi vengono rappresentati tramite l'arco indiretto "-?-". Successivamente, vengono generate tutte le possibili combinazioni tra gli elementi del primo e del secondo nodo per completare il nuovo albero convertito. È fondamentale verificare che l'arco da inserire non sia già presente nell'albero, poiché l'input del sistema MASTRO non consente la duplicazione degli archi. Di seguito è riportato il pseudocodice corrispondente alla funzione.

Algorithm 2 create_edge(converter, first_node, second_node, edge, temp)

```
    Input: converter, first_node, second_node, edge, temp
2:  Output: temp
    Get first_node and second_node
4:  Create combinations if more pathways are in the same node        ▷ -?- edge
    if first_node ≠ [''] and second_node ≠ [''] then
6:      Create a matrix of combinations between first_node and second_node
        for each pair in the matrix do
8:          if the combination is not in temp then
                Add the combination to temp
10:         end if
        end for
12: end if
    return temp
```

2.1.1 Identificazione del nodo di maggiore importanza più vicino alla radice

In questa analisi, si considereranno di maggiore importanza i nodi che appaiono più vicini alla radice.

La funzione *update_score(node_dict, edge)* è progettata per calcolare il punteggio associato a ciascun nodo. Questo punteggio viene determinato contando il numero di antenati e discendenti di ogni nodo all'interno della rete. L'algoritmo analizza gli archi (edge) forniti e aggiorna il dizionario dei nodi (node_dict) di conseguenza, assicurando che ogni nodo contenga le informazioni corrette sui propri antenati e discendenti.

Algorithm 3 update_score(node_dict, edge)

```
1:  Input: node_dict, edge
2:  if "->-" in edge then
3:      Get first_node and second_node
4:      if first_node not in node_dict then
5:          Update node_dict with the new node
6:      else
7:          Update the number of descendants (+1)
8:      end if
9:      if second_node not in node_dict then
10:         Update node_dict with the new node
11:     else
12:         Update the number of ancestors (+1)
13:     end if
14: else if ("-/-" or "-?-") in edge then
15:     Get first_node and second_node and add to the dictionary
16: end if
```

Nella funzione *close_to_root_criterio(node_dict)* vengono mantenuti solo i pathways associati ai nodi con il minor numero di antenati. Nel caso in cui due nodi abbiano lo stesso numero di antenati, viene considerato importante il nodo con il maggior numero di discendenti. Una volta ottenuto il dizionario contenente i pathways più rilevanti, è possibile procedere alla creazione di un dizionario filtrato che consente la costruzione di un nuovo albero convertito.

Algorithm 4 close_to_root_criterio(node_dict)

```
Input: node_dict                          ▷ This dictionary stores the score of every node
Output: pathway_info                      ▷ Significant pathway is associate to specific gene
Initialize pathway_info as an empty dictionary
for all (gene, info) in node_dict do
    for all pathway in info['pathways'] do
        if info['ancestor'] < pathway_info[pathway]['min_ancestor'] then
            Update pathway_info[pathway] with current info and gene
        else if info['ancestor'] == pathway_info[pathway]['min_ancestor'] then
            if info['descentant'] > pathway_info[pathway]['max_descentant'] then
                Update pathway_info[pathway] with current descentant and gene
            end if
        end if
    end for
end for
return pathway_info
```

2.1.2 Identificazione del nodo di maggiore importanza in una posizione interna casuale

Prima di analizzare la funzione relativa a questo criterio di importanza, è opportuno definire i diversi tipi di nodi che possono comparire negli alberi dei dataset in ingresso. I nodi sono classificabili come segue:

- **Nodo radice**: un nodo privo di antenati, ovvero il nodo principale da cui partono tutti gli altri nodi dell'albero.
- **Nodo foglia**: un nodo privo di discendenti, che rappresenta una delle estremità dell'albero.
- **Nodo interno**: un nodo che possiede almeno un discendente, ovvero un nodo situato all'interno dell'albero che non è né la radice né una foglia.

In questo caso la funzione *update_score(node_dict, edge)* calcola lo score in modo da calcolare il punteggio in base alla classificazione del nodo (radice, foglia o interno).

Nella funzione *rnd_int_node_criterio(node_dict)*, vengono conservati esclusivamente i pathways associati a nodi di tipo interno. Qualora non siano presenti nodi interni, la funzione seleziona casualmente un nodo radice o un nodo foglia. È importante notare che questa scelta non è completamente casuale, ma segue l'ordine di apparizione dei nodi fornito dal dataset.

Algorithm 5 update_score(node_dict, edge)

Input: node_dict, edge
if "->-" in edge **then**
    Get *first_node* and *second_node*
    if first_node not in node_dict **then**
        Set the node as root node
    else
        Set the node as internal node
    end if
    if second_node not in node_dict **then**
        Set the node as leaf node
    else
        Set the node as internal node
    end if
else if ("-/-" or "-?-") in edge **then**
    Get *first_node* and *second_node* and add to the dictionary
end if

Algorithm 6 rnd_int_node_criterio(node_dict)

Input: node_dict ▷ This dictionary stores the score of every node
Output: pathway_info ▷ Significant pathway is associate to specific gene
Initialize *pathway_info* as empty dictionary
for each gene, info in node_dict **do**
    for each pathway in info['pathways'] **do**
        if info['ancestor'] == 1 **and** info['descendant'] == 1 **then**
            Set importance to 1 ▷ Internal node
        else if (info['ancestor'] == 0 **or** info['descendant'] == 0) **and** pathway **not in** pathway_info **then**
            Set importance to 0 ▷ Root or leaf node
        end if
    end for
end for

2.1.3 Analisi comparativa degli alberi prima e dopo la conversione

Di seguito vengono illustrati alcuni alberi che sono stati convertiti per evidenziare le differenze tra le due diverse conversioni e confrontarle con l'albero originale. Questa rappresentazione visiva permette di osservare chiaramente le variazioni introdotte da ciascun metodo di conversione e come questi impattino sulla struttura dell'albero di partenza.

E' presentato un esempio relativo ad un albero del dataset AML.

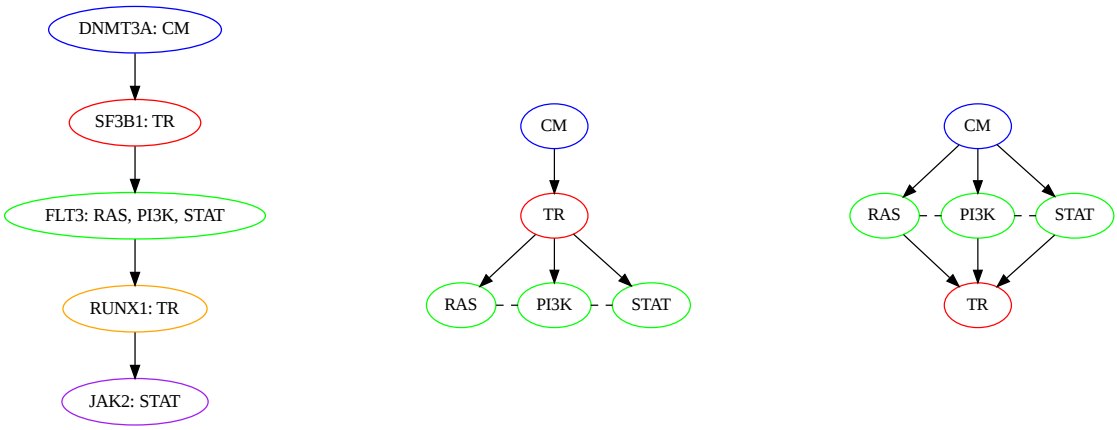

Figura 2.1: Il primo rappresenta l'albero filogenetico di partenza in cui si può osservare la mappatura gene/pathways, il secondo è la conversione "close_to_root", il terzo è la conversione "rnd_int_node"

Albero originale:

```
FLT3->-JAK2 SF3B1->-FLT3 DNMT3A->-FLT3 FLT3->-RUNX1 SF3B1->-JAK2
DNMT3A->-JAK2 RUNX1->-JAK2 DNMT3A->-SF3B1 SF3B1->-RUNX1 DNMT3A->-RUNX1
```

Albero convertito con la convenzione "close_to_root":

```
RAS-?-PI3K RAS-?-STAT PI3K-?-STAT TR->-RAS TR->-PI3K TR->-STAT CM->-RAS
CM->-PI3K CM->-STAT CM->-TR
```

Albero convertito con la convenzione "rnd_int_node":

```
RAS-?-PI3K RAS-?-STAT PI3K-?-STAT CM->-RAS CM->-PI3K CM->-STAT RAS->-TR
PI3K->-TR STAT->-TR CM->-TR
```

Qui è mostrato un esempio del dataset NSCLC.

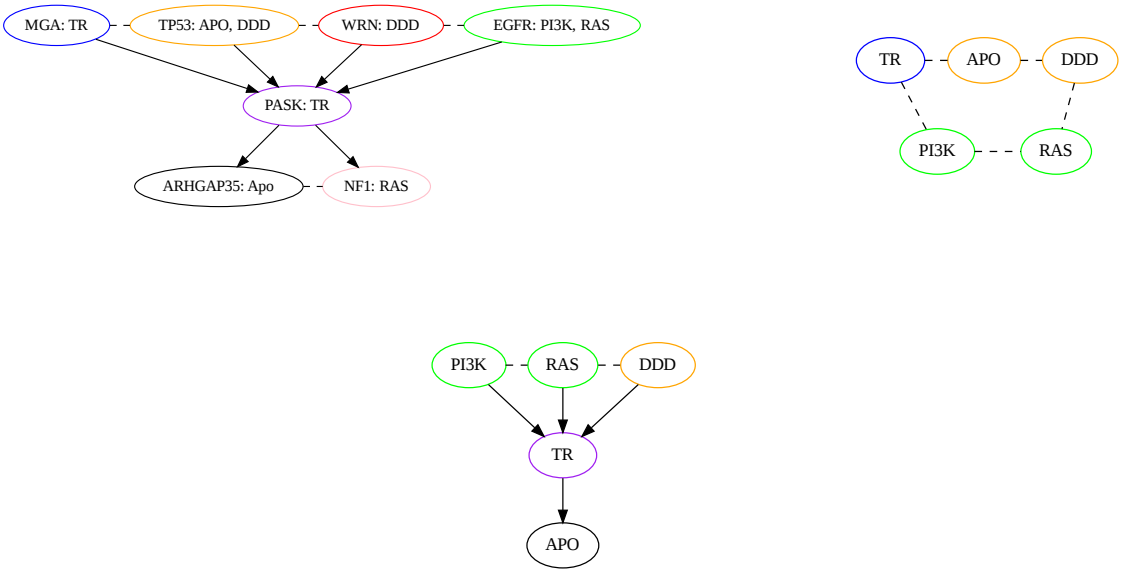

Figura 2.2: Il primo rappresenta l'albero filogenetico di partenza in cui si può osservare la mappatura gene/pathways, il secondo è la conversione "close_to_root", il terzo è la conversione "rnd_int_node"

Albero originale:

```
EGFR->-PASK WRN->-PASK WRN->-ARHGAP35 MGA->-PASK EGFR-?-TP53 WRN->-
NF1 MGA->-NF1 TP53-?-WRN MGA-?-TP53 MGA-?-WRN PASK->-ARHGAP35 MGA-
>-ARHGAP35 PASK->-NF1 ARHGAP35-/-NF1 EGFR->-ARHGAP35 EGFR->-NF1
EGFR-?-MGA TP53->-NF1 EGFR-?-WRN TP53->-ARHGAP35 TP53->-PASK
```

Albero convertito con la convenzione "close_to_root":

```
PI3K-?-RAS PI3K-?-APO RAS-?-APO APO-?-DDD TR-?-APO TR-?-DDD PI3K-?-TR
RAS-?-TR PI3K-?-DDD RAS-?-DDD
```

Albero convertito con la convenzione "rnd_int_node":

```
PI3K-?-RAS PI3K->-TR RAS->-TR DDD->-TR DDD->-APO TR->-APO PI3K->-APO
RAS->-APO PI3K-?-DDD RAS-?-DDD
```

Capitolo 3

Analisi e confronto dei dati e dei risultati

In questo capitolo, vengono discussi i risultati delle conversioni dei dataset citati in precedenza, utilizzando i due metodi di conversione precedentemente descritti. Le analisi sul numero dei pattern e sulle traiettorie sono state effettuate grazie all'utilizzo di MASTRO. Le analisi sono state eseguite con i parametri di input predefiniti, ossia un supporto minimo delle traiettorie pari a 2 e una permutazione indipendente. Per quanto riguarda l'analisi delle traiettorie restituite in output da MASTRO, sono state selezionate tutte, senza applicare alcun tipo di filtraggio. Tutti i risultati e i grafici a cui si farà riferimento sono presenti nella pagina GitHub citata in precedenza.

3.1 Dataset AML

E' stata condotta un'analisi comparativa tra i risultati ottenuti dalla conversione e gli alberi originali di partenza. Tale analisi si è focalizzata, prima di tutto, sulla distribuzione media della profondità e sulla distribuzione media dei nodi. I risultati hanno evidenziato una riduzione nel numero complessivo dei nodi a seguito della conversione. Questo esito era atteso, poiché un maggior numero di geni viene ora mappato sugli stessi pathways; di conseguenza, non potendo esistere nodi con lo stesso pathway in un albero, alcuni nodi sono stati eliminati in conformità con le due regole di conversione stabilite.

La figura 3.1 mostra più in dettaglio la distribuzione del numero di nodi considerando gli alberi originali (a sinistra), gli alberi dopo la conversione in cui il nodo più significativo è quello che appare più vicino alla radice (al centro), e gli alberi dopo la conversione dove il nodo più significativo è un nodo interno casuale.

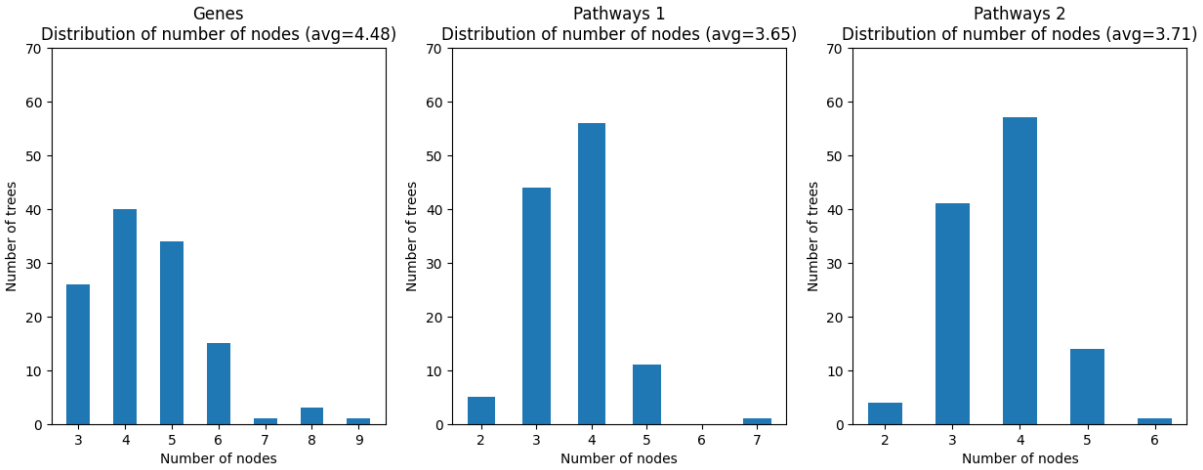

Figura 3.1: Analisi del numero di nodi prima e dopo la conversione

Inoltre, poiché la conversione ha comportato la rimozione di alcuni nodi, anche la profondità media degli alberi è risultata ridotta rispetto a quella degli alberi originali.

Nella Figura 3.2, sono rappresentate: la profondità degli alberi originali (a destra), gli alberi dopo la conversione, in cui il nodo più significativo è quello più vicino alla radice (al centro), e gli alberi convertiti in cui il nodo più significativo è un nodo interno casuale (a sinistra).

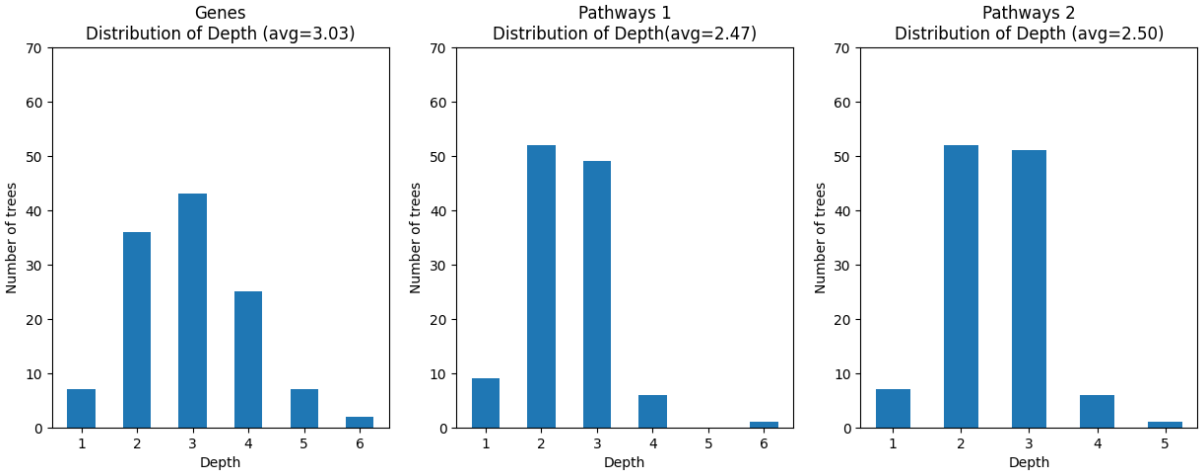

Figura 3.2: Analisi della profondità prima e dopo la conversione

Successivamente, è stata effettuata un'analisi dei risultati utilizzando l'algoritmo MASTRO, con l'obiettivo di confrontare il numero di pattern identificati dopo la conversione con la media delle traiettorie osservate. I dati hanno rivelato una riduzione nel numero complessivo di pattern individuati dall'algoritmo, accompagnata da un aumento medio delle traiettorie rilevate. Questo fenomeno può essere attribuito al fatto che un numero maggiore di geni viene ora mappato su pathways simili, il che comporta una diminuzione del numero totale di pattern distintivi, ma un incremento della frequenza dei pattern più comuni.

La Figura 3.3 evidenzia i cambiamenti tra gli alberi originali (a destra), gli alberi dopo la conversione in cui il nodo più significativo è quello più vicino alla radice (al centro), e gli alberi convertiti in cui il nodo più significativo è un nodo interno casuale (a sinistra).

	Gene	Pathways 1	Pathways 2
Nr. Pattern	137.0	92.0	102.0
Media Traiettorie	3.7664233576642334	9.380434782608695	8.607843137254902

Figura 3.3: Nr. pattern e media delle traiettorie

3.1.1 Analisi delle traiettorie più frequenti

In questa sezione viene svolta un'analisi dettagliata delle traiettorie con il maggior numero di supporti, valutando sia le similarità che le differenze tra i dataset analizzati. Le traiettorie rappresentano le relazioni tra nodi in reti biologiche, individuate mediante lo strumento MASTRO. I numeri tra parentesi indicano la quantità di supporti per ciascuna traiettoria, mentre il simbolo *g* identifica la radice dell'albero. In particolare, verranno confrontati il dataset originale e due varianti ottenute tramite differenti criteri di conversione.

In questi risultati, traiettorie del tipo *A->-B* e *B->-A* indicano che i pathways A e B appaiono nello stesso nodo.

Dataset originale

Il dataset originale, riportato di seguito, presenta tre traiettorie che includono tre geni noti nel contesto biologico: DNMT3A, NPM1 e FLT3.

```
DNMT3A->-NPM1,g->-DNMT3A,g->-NPM1 (17)
DNMT3A->-FLT3,g->-DNMT3A,g->-FLT3 (13)
NPM1->-FLT3,g->-NPM1,g->-FLT3 (13)
```

Dataset con convezione vicino alla radice

In questa variante del dataset, viene adottata una convenzione che posiziona il nodo biologicamente più rilevante il più vicino possibile alla radice. Tale approccio permette di evidenziare l'importanza di alcuni geni chiave nel contesto biologico analizzato.

```
PI3K->-STAT,STAT->-PI3K,g->-PI3K,g->-STAT (50)
CM->-RAS,g->-CM,g->-RAS (42)
RAS->-STAT,STAT->-RAS,g->-RAS,g->-STAT (41)
```

In questo caso, le traiettorie mostrano come i geni PI3K, STAT e RAS siano posizionati in prossimità della radice, conferendo loro una centralità maggiore rispetto agli altri. I valori di supporto sono significativamente più alti rispetto al dataset originale, indicando una maggiore rilevanza di queste traiettorie nel contesto analizzato.

Dataset con convezione nodo interno casuale

In questa seconda variante di conversione, invece, viene adottata una convenzione che assegna l'importanza a un nodo interno in maniera casuale, mantenendo comunque una struttura ad albero coerente con i dati originali.

```
PI3K->-STAT,STAT->-PI3K,g->-PI3K,g->-STAT (50)
CM->-RAS,g->-CM,g->-RAS (40)
RAS->-STAT,STAT->-RAS,g->-RAS,g->-STAT (37)
```

I risultati in questo caso mantengono la stessa struttura delle traiettorie rispetto alla conversione precedente, ma i supporti risultano leggermente inferiori, segnalando una variazione minima nella rilevanza delle traiettorie.

Nonostante le diverse convenzioni utilizzate per la conversione del dataset, i risultati principali sono in gran parte consistenti. Le traiettorie con maggior supporto rimangono sostanzialmente invariate sia per struttura che per frequenza. In particolare, le traiettorie che coinvolgono i geni PI3K, STAT e RAS mantengono un supporto elevato, sottolineando la centralità di questi geni nei pathway considerati.

Una differenza significativa tra i dataset riguarda la mappatura dei pathway associati ai diversi geni. Ad esempio, nel dataset originale, il gene FLT3 appare nelle traiettorie in associazione con DNMT3A e NPM1, ma nei dataset convertiti emerge che FLT3 è collegato a pathway cruciali come RAS, PI3K e STAT. Questo suggerisce che, in termini di centralità e importanza biologica, FLT3 possa avere un ruolo più rilevante di quanto indicato nel dataset originale.

D'altro canto, è interessante notare l'assenza del gene NPM1 nelle traiettorie principali nel dataset convertito, nonostante la sua presenza nel dataset originale. Questo gene è noto per essere associato al pathway Apoptosis, ma sembra non essere rappresentato tra i pathway più diffusi nei dataset convertiti. Questa assenza può indicare una minore centralità di NPM1 nelle traiettorie più frequentemente supportate o un ruolo meno rilevante nel contesto dei pathway principali esplorati

3.2 Dataset NSCLC

Il dataset NSCLC è caratterizzato da alberi che presentano al massimo 5 nodi. Questa limitazione deriva dall'abbondanza di alterazioni osservate, che impedisce un ordinamento affidabile dei nodi stessi. Analogamente a quanto effettuato per il dataset AML, si è inizialmente analizzata la distribuzione media del numero di nodi. I grafici risultanti dalla conversione degli alberi evidenziano una diminuzione del numero medio di nodi per albero, in linea con le previsioni. Infatti, la maggior parte degli alberi presenta ora 2 nodi nella prima conversione. È interessante notare che, scegliendo come nodo più rilevante un nodo interno casuale, rimane una piccola percentuale di alberi con 5 nodi.

La figura 3.4 mostra più in dettaglio la distribuzione del numero di nodi considerando gli alberi originali (a sinistra), gli alberi dopo la conversione in cui il nodo più significativo è quello che appare più vicino alla radice (al centro), e gli alberi dopo la conversione dove il nodo più significativo è un nodo interno casuale.

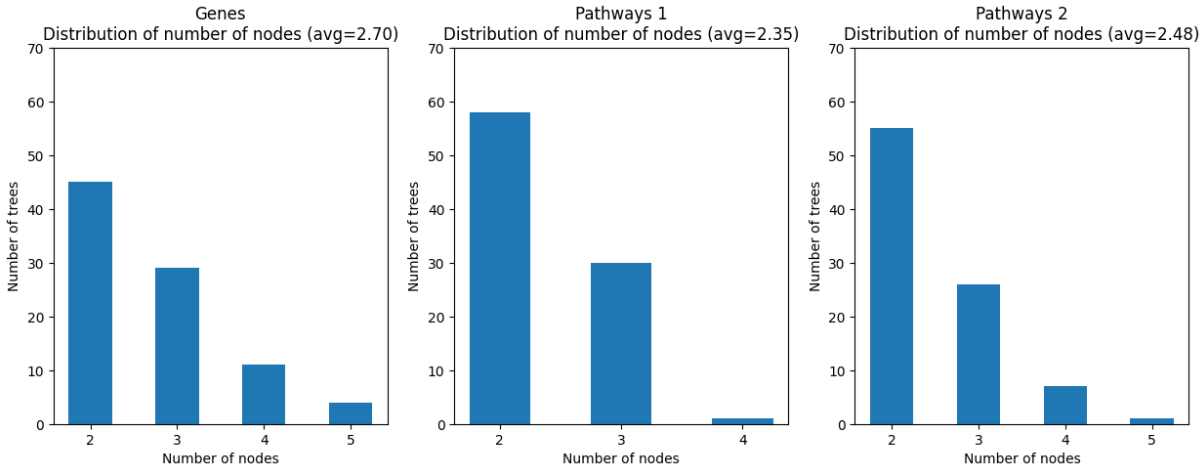

Figura 3.4: Analisi del numero di nodi prima e dopo la conversione

Per quanto riguarda la distribuzione della profondità, poiché, come precedentemente accennato, l'ordinamento dei nodi risulta impossibile da determinare in questo dataset, i risultati delle analisi successive alla conversione mostrano un aumento degli alberi con un solo livello di profondità.

Nella Figura 3.5, sono rappresentate: la profondità degli alberi originali (a destra), gli alberi dopo la conversione, in cui il nodo più significativo è quello più vicino alla radice (al centro), e gli alberi convertiti in cui il nodo più significativo è un nodo interno casuale (a sinistra).

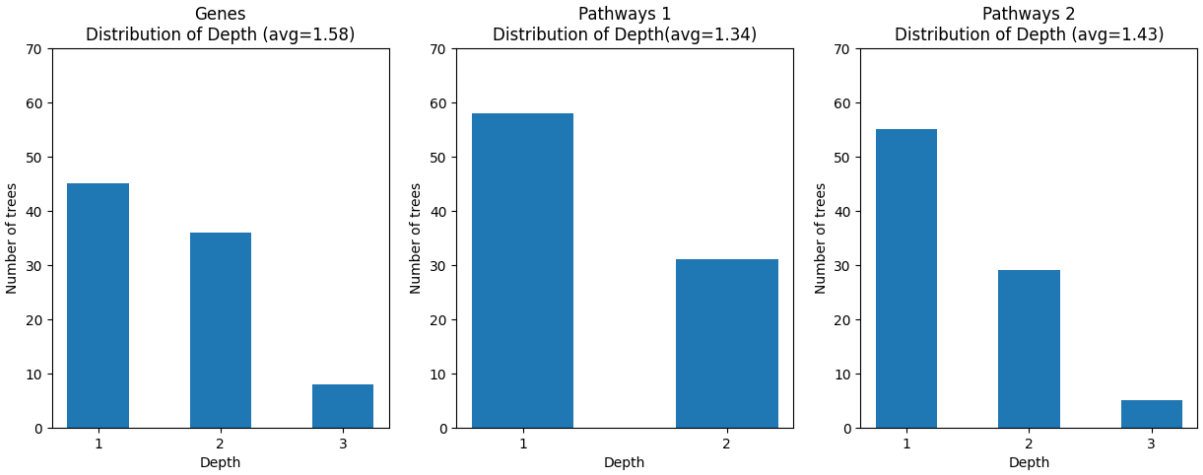

Figura 3.5: Analisi della profondità prima e dopo la conversione

Successivamente, è stato osservato un notevole aumento nel numero di pattern rispetto agli alberi originali, verosimilmente dovuto alla mappatura di pathways simili associati a geni diversi. Questo fenomeno ha comportato un raddoppio dei pattern individuati dal sistema MASTRO e un incremento significativo del numero di occorrenze di tali pattern all'interno del dataset.

La Figura 3.6 evidenzia i cambiamenti tra gli alberi originali (a destra), gli alberi dopo la conversione in cui il nodo più significativo è quello più vicino alla radice (al centro), e gli alberi convertiti in cui il nodo più significativo è un nodo interno casuale (a sinistra).

	Gene	Pathways 1	Pathways 2
Nr. Pattern	123.0	279.0	278.0
Media Traiettorie	2.886178861788618	7.232974910394265	6.190647482014389

Figura 3.6: Nr. pattern e media delle traiettorie

3.2.1 Analisi delle traiettorie più frequenti

In questa sezione vengono esaminate le traiettorie con il maggior numero di supporti, analizzando le loro similarità e differenze attraverso tre dataset: quello originale e due versioni convertite. Il numero tra parentesi indica i supporti individuati, mentre *g* rappresenta la radice dell'albero.

In questi risultati, traiettorie del tipo *A->-B* e *B->-A* indicano che i pathways A e B appaiono nello stesso nodo.

Dataset originale

Il dataset originale presenta traiettorie che coinvolgono i geni SOX2, TP53 e PIK3CA, con amplificazione genica (amp) associata a SOX2 e PIK3CA. Questi geni mappano sui pathway biologici WNT e NOTCH (SOX2), Apoptosis e DDD (TP53), e PI3K (PIK3CA).

```
SOX2amp->-TP53,TP53->-SOX2amp,g->-SOX2amp,g->-TP53 (12)
PIK3CAamp->-SOX2amp,SOX2amp->-PIK3CAamp,g->-PIK3CAamp,g->-SOX2amp (12)
PIK3CAamp->-TP53,PIK3CAamp->-SOX2amp,TP53->-PIK3CAamp,TP53->-
SOX2amp,SOX2amp->-PIK3CAamp,SOX2amp->-TP53,g->-PIK3CAamp,g->-TP53,g-
>-SOX2amp (10)
```

Le traiettorie evidenziano relazioni complesse tra i tre geni, con frequenze simili, suggerendo un'interazione forte tra i pathway WNT, Apoptosis, PI3K e DDD. Le frequenze di supporto sono simili, il che indica che queste traiettorie sono comunemente riscontrate nei dati.

Dataset con nodo più importante vicino alla radice:

In questa versione del dataset, i nodi biologicamente più rilevanti sono posizionati più vicino alla radice. Si osservano cambiamenti significativi, con il pathway Apoptosis e DDD che emergono come centrali nelle traiettorie.

```
Apoptosis->-DDD,DDD->-Apoptosis,g->-Apoptosis,g->-DDD (51)
DDD->-Apoptosis,Apoptosis->-DDD,g->-DDD,g->-Apoptosis (51)
RAS->-Apoptosis,Apoptosis->-RAS,g->-RAS,g->-Apoptosis (22)
```

Rispetto al dataset originale, emerge una forte centralità del pathway Apoptosis, mappato da TP53, mentre il gene SOX2 e il pathway WNT non sono più presenti. Il supporto per Apoptosis e DDD è molto elevato (51), suggerendo una maggiore rilevanza di questi pathway nelle traiettorie più supportate.

Dataset con nodo interno casuale più importante:

In questa versione, i nodi centrali sono selezionati casualmente. Le traiettorie mantengono la stessa struttura, ma con supporti leggermente inferiori.

```
Apoptosis->-DDD,DDD->-Apoptosis,g->-Apoptosis,g->-DDD (43)
DDD->-Apoptosis,Apoptosis->-DDD,g->-DDD,g->-Apoptosis (43)
RAS->-Apoptosis,Apoptosis->-RAS,g->-RAS,g->-Apoptosis (23)
```

Anche qui, Apoptosis e DDD rimangono centrali nelle traiettorie, con supporti leggermente inferiori rispetto alla conversione precedente. La stabilità della struttura suggerisce che i pathway legati a Apoptosis e DDD, associati a TP53, continuano a dominare la rete biologica considerata.

Nonostante le diverse convenzioni utilizzate, i risultati mostrano una notevole coerenza. Le traiettorie con maggior supporto coinvolgono principalmente i pathway Apoptosis e DDD, mappati dal gene TP53. Nella conversione con il nodo più importante vicino alla radice, i supporti per Apoptosis e DDD aumentano, confermando il loro ruolo dominante.

In contrasto, nel dataset originale i geni SOX2 e PIK3CA, associati rispettivamente ai pathway WNT, NOTCH e PI3K, erano più centrali, ma scompaiono nelle versioni convertite, lasciando spazio a traiettorie dominate da Apoptosis e DDD.

Questa differenza riflette come la centralità di un gene o pathway possa variare a seconda della convenzione utilizzata, suggerendo che nel contesto delle reti biologiche, il gene TP53 risulti maggiormente implicato nelle traiettorie più significative, mentre SOX2 e PIK3CA potrebbero avere un ruolo meno dominante rispetto a quanto indicato inizialmente.

Capitolo 4

Conclusioni

In questa tesi è stato affrontato il problema dell'analisi delle traiettorie evolutive tumorali a livello di pathway utilizzando alberi filogenetici e l'algoritmo MASTRO. L'obiettivo principale era sviluppare un metodo per convertire alberi genetici in alberi basati su pathway, al fine di offrire una prospettiva diversa sull'evoluzione dei tumori. I risultati ottenuti dimostrano che le conversioni realizzate, secondo i criteri di prossimità alla radice e nodi interni casuali, hanno prodotto una riduzione complessiva della complessità degli alberi, con una diminuzione del numero di nodi e della profondità degli alberi convertiti.

L'analisi comparativa dei dataset AML e NSCLC ha rivelato che, nonostante la semplificazione degli alberi, le traiettorie più significative sono rimaste conservate, consentendo comunque una rappresentazione accurata delle alterazioni genetiche. In particolare, si è osservato un aumento delle traiettorie comuni nei dataset convertiti, evidenziando come la mappatura gene-pathway possa ridurre la varietà di traiettorie mantenendo una rappresentazione coerente dell'evoluzione tumorale.

Questi risultati dimostrano che l'approccio proposto è efficace nel ridurre la complessità dei dati senza compromettere la qualità dell'analisi delle traiettorie. Tuttavia, ulteriori sviluppi potrebbero migliorare ulteriormente l'accuratezza delle conversioni, ad esempio integrando criteri di rilevanza più complessi o estendendo l'analisi ad altre tipologie di tumori. Nel complesso, il lavoro svolto apre nuove possibilità per l'interpretazione delle dinamiche evolutive nei tumori e offre una base per futuri studi sull'argomento.

Ringraziamenti

Desidero dedicare questo spazio a tutte le persone che mi hanno sostenuto nel mio percorso universitario e a coloro che, in un modo o nell'altro, sono sempre stati al mio fianco.

Un sincero ringraziamento va a mia mamma, mio papà, Diego e mia nonna Anna per aver sopportato ogni mia giornata negativa e per avermi costantemente sostenuto durante questi anni di studio.

Grazie a Federico per essere sempre stato al mio fianco, sia nei momenti felici che in quelli difficili.

Grazie ad Alberto per essere diventato come il fratello maggiore che non ho mai avuto. Grazie Riccardo ed Asya che sono sempre pronti ad ascoltarmi e a sostenermi ogni volta che ne ho avuto bisogno. E grazie a tutti e tre per avermi ospitato così tanto in estate, facendomi dormire più da voi che a casa mia.

Grazie di cuore a tutti gli amici che ho incontrato durante questi tre anni e a quelli che mi accompagnano da prima dell'inizio del mio percorso universitario.

Bibliografia

[1] Mattia Bastianello. *genes_pathways_converter*. Ver. 1.0. url: https://github.com/Mattichs/genes_pathways_converter.

[2] Leonardo Pellegrina e Fabio Vandin. «Discovering significant evolutionary trajectories in cancer phylogenies». In: *Bioinformatics* 38.Supplement$_2$ (set. 2022), pp. ii49–ii55. issn: 1367-4803. doi: 10.1093/bioinformatics/btac467. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/Supplement_2/ii49/49886558/btac467.pdf. url: https://doi.org/10.1093/bioinformatics/btac467.