Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea in Ingegneria Informatica

Identification of Significant Evolutionary Trajectories in Cancer

Relatore:
Prof. Pellergrina Leonardo
Correlatore:
Prof. Vandin Fabio

Laureando:
Alessandro Viespoli
2009659

Anno Accademico 2022/2023
Data di Laurea 25/09/2023

Abstract

I tumori sono il risultato di processi di evoluzione somatica che col passare del tempo portano, all'interno del stesso tumore, allo sviluppo di zone geneticamente distinte. Negli ultimi anni diverse tecniche di campionamento e sequenziamento hanno permesso l'analisi di singole o molteplici regioni tumorali, evidenziando che durante lo sviluppo di un tumore esistono caratteristiche comuni condivise fra tumori dello stesso genere in pazienti diversi. I tumori maligni, noti anche come cancro, sono una delle maggiori cause di morte nel mondo, quindi l'identificazione di pattern comuni nella evoluzione tumorale permetterebbe di creare terapie e cure efficaci.

Questo elaborato di tesi presenta i risultati ottenuti dallo studio e analisi di alberi filogenetici di pazienti con cancro al seno tramite l'algoritmo MASTRO, il quale permette di identificare traiettorie di mutazioni frequenti in un insieme di pazienti. Il dataset considerato è composto da 37809 alberi filogenetici distribuiti tra 1315 pazienti affetti da cancro al seno.

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Nell'ambito oncologico, una delle sfide più impegnative è quella di studiare e predire il comportamento dei tumori. I tumori sono il risultato di processi di evoluzione somatica, per questo motivo negli ultimi anni diverse tecniche di campionamento e analisi del DNA hanno svolto un ruolo chiave verso la lotta al cancro.
Col passare del tempo i tumori crescono, si espandono, facendo emergere nuove sottopopolazioni di cellule tumorali con alterazioni genomiche distinte, dando vita alla eterogeneità intra-tumorale. Proprio per questo motivo, molti progressi significativi sono stati raggiunti nel sequenziamento cellulare multi regione ed a singola cellula, permettendo una migliore rappresentazione dell'architettura clonale dei tumori.
I dati raccolti hanno mostrato che, mentre è presente una componente stocastica intrinseca nell'evoluzione dei tumori, esistono alcune caratteristiche comuni che sono condivise nella progressione di certe tipologie di tumore. Il rilevamento di queste caratteristiche comuni è cruciale per predire il comportamento dei tumori e per sviluppare terapie mirate ed efficaci.
Negli ultimi anni, diversi metodi computazionali sono stati sviluppati con l'obbiettivo di inferire la composizione subclonale dei tumori oppure di inferire la storia evolutiva di questi tramite i dati ottenuti dal sequenziamento tumorale; questi metodi producono in output un albero filogenetico (*phylogenetic tree*), che rappresenta uno dei possibili ordini in cui queste alterazioni genomiche vengono osservate.
In questo elaborato vedremo l'algoritmo MAximal tumor treeS TRajectOries (MASTRO) applicato a un dataset di 37809 alberi filogenetici distribuiti tra 1315 pazienti affetti da cancro al seno.

Capitolo 1

Cancerogenesi

Il cancro è una della maggiori cause di morte nel mondo e, nonostante gli enormi sforzi e risorse spese, l'eliminazione o il controllo di questa malattia rimane una sfida significativa.
Il processo che porta allo sviluppo del cancro, detto cancerogenesi, può compiersi in tempi brevi o nell'arco di anni o addirittura di decenni. In generale il processo prende il via da una cellula progenitrice che subisce un'alterazione genetica; questa alterazione poi verrà trasmessa ai figli che ne deriveranno dalla sua moltiplicazione (cloni).
Per capire questo meccanismo, vengono qui introdotti il concetto di DNA, come nasce un tumore e il funzionamento alla base dell'evoluzione tumorale.

1.1 DNA

Le informazioni genetiche risiedono nella sequenza dell'acido desossiribonucleico (DNA); il DNA è contenuto nel nucleo delle cellule sotto forma di cromosomi. Dal punto di vista chimico, il DNA è un polimero organico, ovvero una macromolecola composta da uno o più atomi di carbonio, a doppia catena la cui unità base è chiamata nucleotide.
I nucleotidi sono formati da tre componenti fondamentali:

1. Un gruppo fosfato (PO_4^{2-})
2. Uno zucchero pentoso, il deossiribosio
3. Una base azotata

A loro volta le basi azotate che entrano nella formazione del nucleotide sono quattro:

- Adenina [A]
- Timina [T]
- Citosina[C]
- Guanina [G]

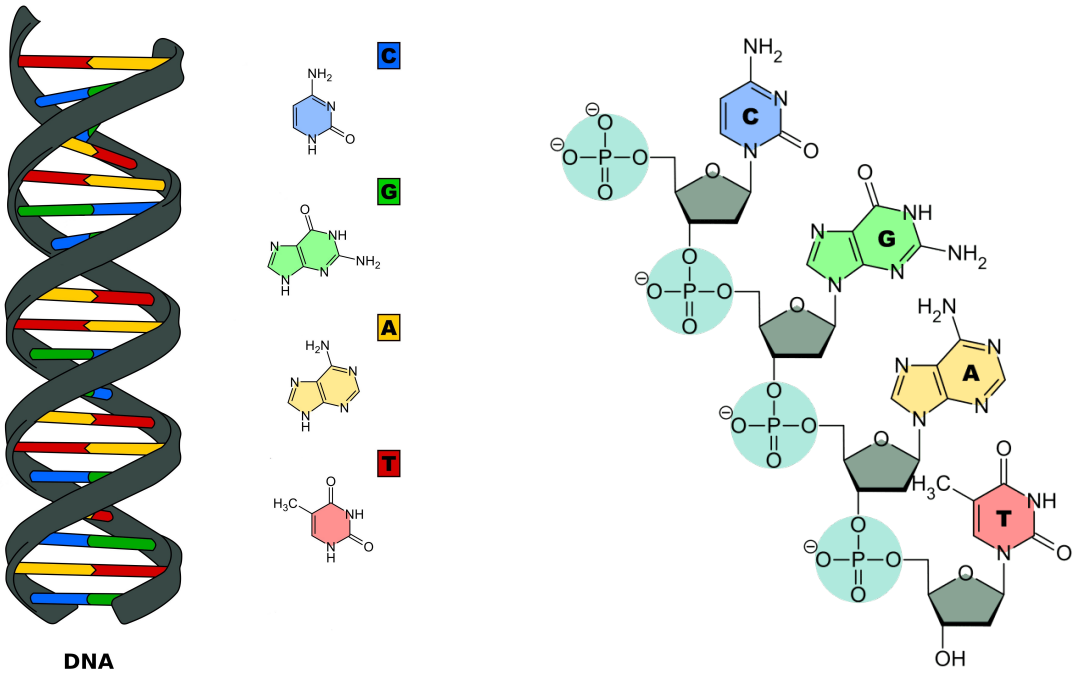


Figura 1.1: Struttura a doppia elica DNA [14]

Quindi il DNA è formato dalla successione di nucleotidi e le 2 eliche caratteristiche sono formate da due catene di nucleotidi complementari, unite tra loro tramite legami covalenti tra le diverse basi azotate, in particolare i legami avvengono tra A-T e C-G.

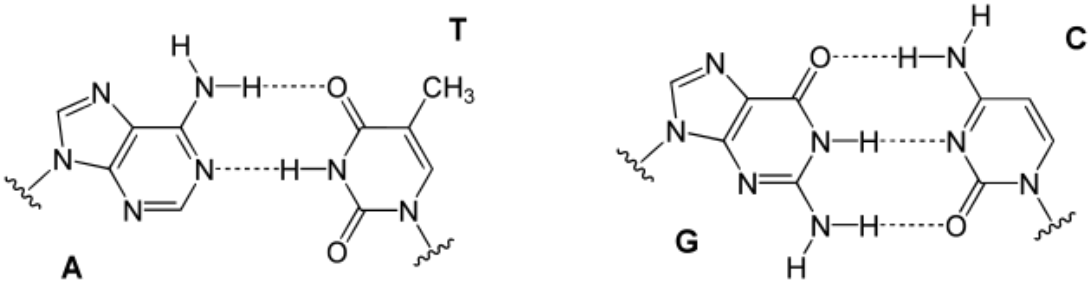


Figura 1.2: Legami covalenti tra basi azotate [15]

La sequenza di queste lettere nelle sue combinazioni determina come un organismo vivente è fatto e quali sono le sue caratteristiche.

1.2 Tumore

1.2.1 Gene

Le unità di base dell'informazione genetica sono i geni. I geni sono delle porzioni di DNA e hanno il compito fondamentale di codificare le proteine, i reali effettori delle funzioni biologiche e i determinanti delle caratteristiche fenotipiche, ovvero le sfumature in cui il nostro corpo si mostra.
I singoli geni non funzionano in modo autonomo, ma dipendono da altre componenti per la replicazione. La totalità di tutti i geni di un organismo forma il genoma.

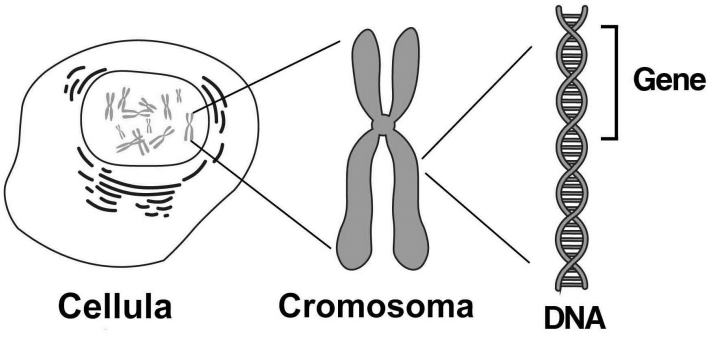


Figura 1.3: Organizzazione cellula, cromosoma, DNA, gene [16]

1.2.2 Mutazioni Genetiche

Normalmente le cellule si moltiplicano seguendo regole e ritmi abbastanza precisi; queste regole che determinano le caratteristiche e la frequenza della proliferazione sono scritte nei geni, quindi regolate dalle proteine che questi geni producono.
Questo equilibrio può essere messo a rischio quando avvengono delle mutazioni genetiche, ovvero una modifica nella sequenza nucleotidica dei geni. Le mutazioni genetiche possono essere causate da diversi fattori: fattori ambientali, fattori genetici, stile di vita.

Quando avviene una mutazione, esiste una possibilità nella quale le proteine, che favoriscono la moltiplicazione cellulare, stimolino la crescita di cellule tumorali se alterate o se non bloccate da altre proteine che normalmente inibiscono tale processo. Un tumore, infatti, si sviluppa quando le cellule crescono fuori controllo, tuttavia, non basta una singola mutazione affinchè nasca un tumore; il processo di formazione di un tumore nasce dalle accomulazioni di più alterazioni genetiche in un arco di tempo piuttosto lungo. La tendenza dei tumori a diventare sempre più aggressivi nel tempo è denominata progressione tumorale. I tumori possono essere classificati in due categorie:

- **Tumori benigni**, masse compatte in espansione che rimangono circoscritte alla sede d'origine.
- **Tumori maligni** (cancro), associati ad una distribuzione del tessuto circostante verso altre sedi.

1.2.3 Etereogenità intra-tumorale

Il processo che porta alla generazione di un tumore prende il via da una cellula progenitrice che nel momento della sua duplicazione acquisisce un'alterazione genetica; questa mutazione verrà trasmessa di conseguenza ai suoi discendenti, detti cloni. I cloni nel momento della duplicazione potrebbero subire a loro volta delle mutazioni che danno origine a simili, ma geneticamente distinguibili, sottopopolazioni dentro lo stesso tumore, chiamate subcloni. Questo meccanismo fa si che all'interno dello stesso tumore coesistano sezioni geneticamente diverse, quindi con possibili trattamenti e cure diverse tra di loro; quindi l'etereogenità tumorale è una delle maggiori sfide verso il trattamento del cancro.

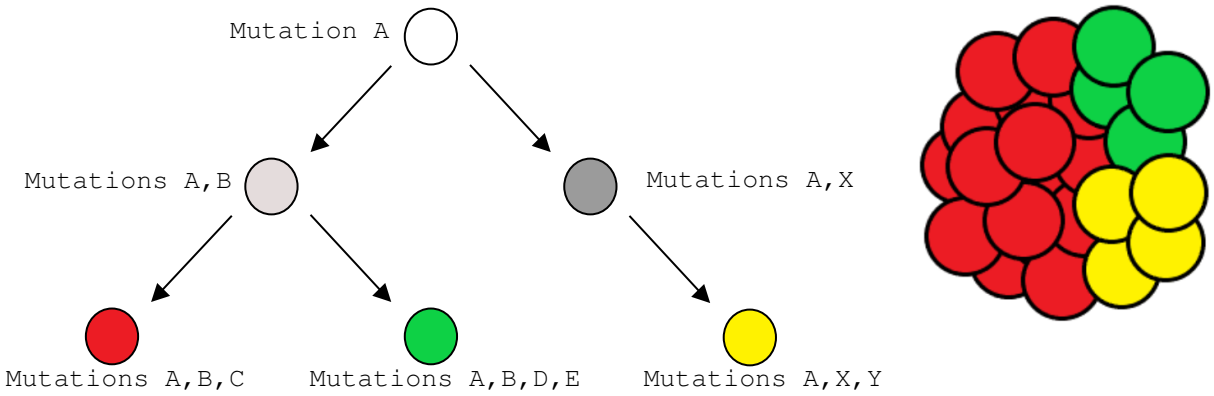


Figura 1.4: Etereogenità intra-tumorale

1.3 Meccanismo di progressione tumorale

I tumori evolvono seguendo un modello di selezione darwiniana poiché, tra i vari subcloni tumorali, vengono a selezionarsi quelli con la maggiore fitness, ovvero che sviluppano le mutazioni che garantiscono maggiore capacità di sopravvivenza. Per questo motivo un tumore che va incontro a recidiva dopo una specifica terapia è molto spesso resistente a quella stessa terapia, poichè sono state selezionate le cellule resistenti.

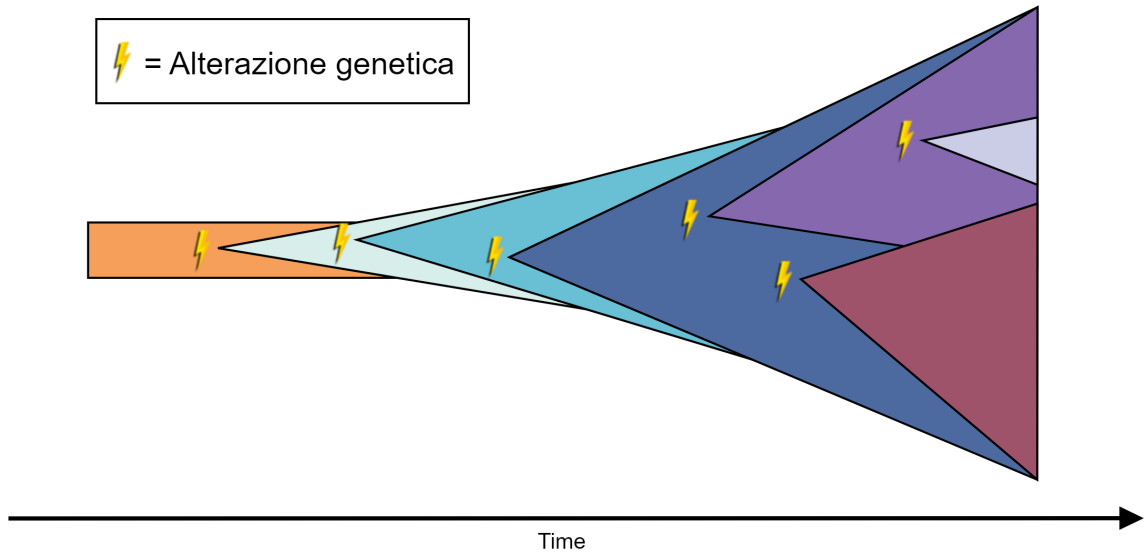


Figura 1.5: Evoluzione di popolazioni subclonali

1.3.1 Modelli

La complessità di un cancro ha origine da semplici concetti di evoluzione che coinvolgono numerosi agenti interagenti (per esempio le bilioni di cellule cancerogene dentro un singolo subclone e il microambiente che circonda il tumore). I principali modelli che entrano in gioco nella progressione tumorale sono:

- **Selezione**. Motivo per cui una "stirpe" è favorita alle altre e produce più discendenti resistenti. La selezione non è sempre operativa, infatti dipende dal contesto ambientale in cui il tumore si trova.
- **Branching evolution**. L'evoluzione è sempre ramificata perchè la divisione cellulare e le mutazioni producono continuamente delle divergenze a livello genomico.
- **Linear evolution**. Il modello di evoluzione lineare prevede che solo una progenie sopravviva; se un singolo clone sopravvive e nel momento del campionamento viene notato allora l'evoluzione appare lineare.
- **Neutral evolution**. Si verifica in assenza di selezione o fitness in una popolazione; può essere vista come l'evoluzione tra gli eventi di selezione. Prima che avvenga una mutazione, il clone evolve in modo neutrale.

- **Punctuated evolution**. I modelli lineari e neutrali assumono come ipotesi che le mutazioni avvengano sequenzialmente e gradualmente col tempo. Tuttavia, recenti studi hanno notato che nei primi stadi tumorali può capitare che ci sia un esplosione di mutazioni genetiche nell'arco di pochissimo tempo definito come punctuated evolution.

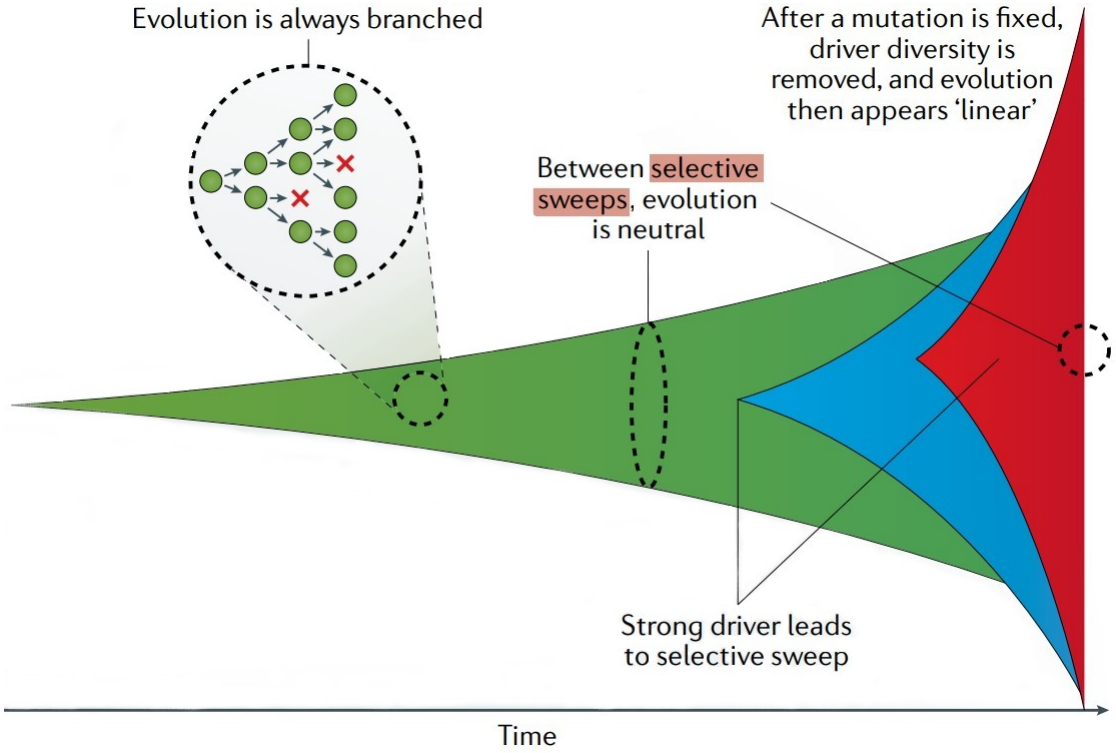


Figura 1.6: Modalità di evoluzione tumorale [6]

Se in piccole sottopopolazioni di un cancro la componente stocastica può essere dominante, in cloni e subcloni più grandi il loro comportamendo risulta essere deterministico. Dunque, l'obbiettivo di MASTRO è di cogliere quest'ultima componente, riuscendo a scoprire le sequenze di mutazioni, ovvero traiettorie, più frequenti per un determinato tipo di tumore.

Capitolo 2

Definizioni Preliminari

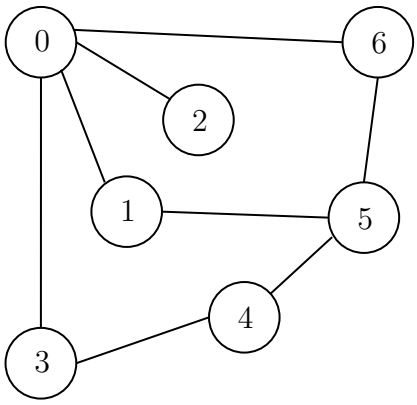
Lo scopo di MASTRO è quello di identificare le più significative traiettorie frequenti in una serie di alberi filogenetici che descrivono l'evoluzione di un gruppo di tumori. Questo capitolo introduce alcune strutture dati e concetti chiave utilizzati dall'algoritmo MASTRO.

2.1 Concetti base

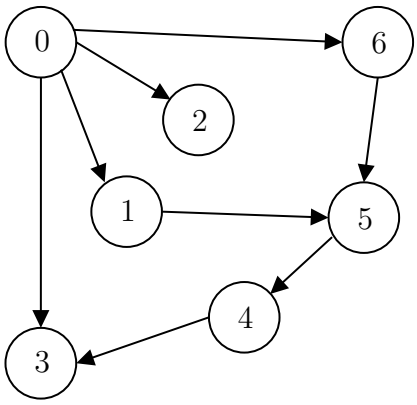
2.1.1 Grafo

Un grafo $G = (V, E)$ è formato da un insieme V di vertici/nodi che possono essere collegati tra loro tramite una collezione E di archi (coppie di vertici).
Un grafo è:

- diretto se ogni arco $(u, v) \in E$ è una coppia ordinata $(u \to v)$
- non diretto se ogni arco $(u, v) \in E$ è una coppia non ordinata $(u - v)$



(a) grafo non diretto



(b) grafo diretto

2.1.2 Sottografo

Un sottogrado $G' = (V', E')$ di un grafo $G = (V, E)$ ha le seguenti caratteristiche:

- $V' \subseteq V$
- $E' \subseteq E$
- $\forall (u, v) \in E' : u, v \in V'$

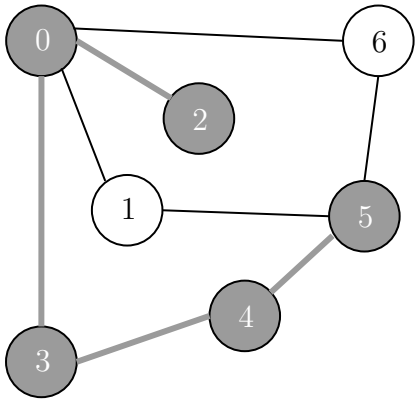


Figura 2.2: Archi e vertici in grigio sono un sottografo della Figura 2.1a

Definizione 2.1 (Sottografo indotto)**.** Un sottografo H di un grafo G si dice indotto se e solo se, per qualsiasi coppia di vertici u e v di H, (u, v) è un arco di H e (u, v) è un arco di G. In altre parole, H è un sottografo indotto di G se ha esattamente gli stessi archi che appaiono in G sullo stesso insieme di vertici.

2.1.3 Albero radicato

Un albero radicato (*rooted tree*) è un grafo $G = (V, E)$ in cui:

- esiste un vertice radice $r \in V$
- $\forall u \in V$, con $u \neq r$, esiste un unico vertice $p(u) \in V$ padre di u
- $E = \{(u, p(u)) : u \in V, u \neq r\}$
- $\forall u \in V$ andando di padre in padre si raggiunge r

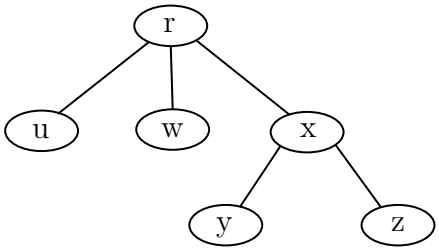


Figura 2.3: Esempio di albero radicato

2.2 Tumor Tree

Un tumor tree $T = (V_T, E_T)$ comprende un insieme V_T di nodi ed un insieme E_T di archi. Ogni nodo di V_T corrisponde ad un clone del tumore e contiene una serie di alterazioni [ad esempio *single-nucleotide variants* (SNVs), *copy number aberrations* (CNA)] provenienti da un insieme A di m alterazioni. Il nodo radice di ogni albero T contiene un insieme vuoto di alterazioni e rappresenta le cellule normali, mentre ogni altro nodo $v \in V_T$ contiene un sottoinsieme non vuoto di alterazioni $\in A$. MASTRO riceve come input un multiinsieme di n *rooted tumor trees* $D = \{T_1, ..., T_n\}$, ricavabili tramite alcuni metodi computazionali per ricostruire la storia filogenetica dei tumori.

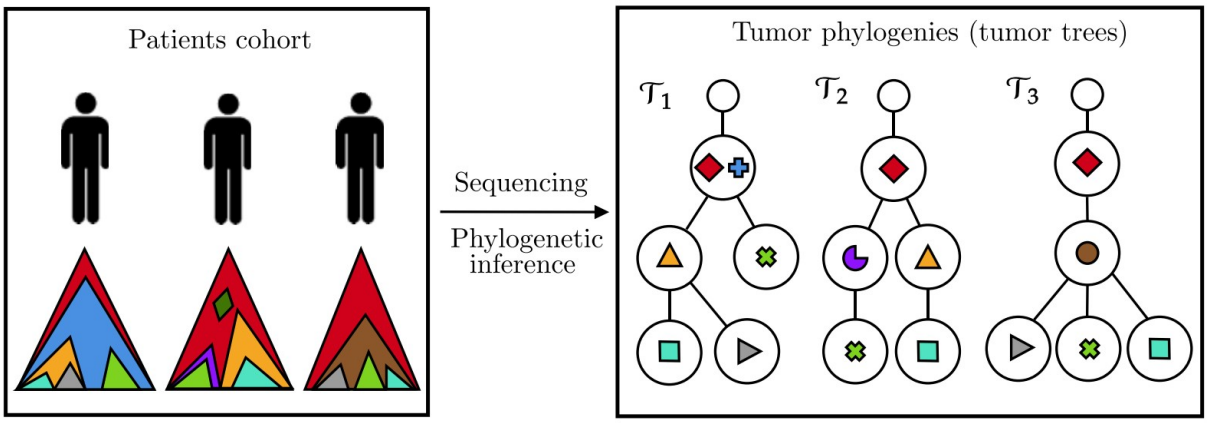


Figura 2.4: Esempio di tre tumor trees [1]

Per un dato nodo il suo insieme di alterazioni ci fornisce le alterazioni che sono apparse per la prima volta nel corrispondende clone del tumore (nodo) ma non nei suoi antenati; dunque per ottenere l'insieme completo di alterazioni per un clone bisogna eseguire l'unione degli insiemi di alterazioni trovati lungo il percorso univoco tra il nodo corrispondente e il nodo radice.

Definizione 2.2 (Antenato di una alterazione)**.** Per ogni tumor tree, una alterazione a, contenuta in v, è un antenato di una alterazione b, contenuta in $w \neq v$, se il nodo v appartiene al percorso da w fino al nodo radice del tumor tree.

2.3 Trajectory

Definizione 2.3 (Trajectory)**.** Una trajectory P è un tumor tree dove $P = (V_P, E_P)$.

Una trajectory P è osservata in un albero T se l'insieme di alterazioni in P è un sottoinsieme delle alterazioni $V_T \in T$ e se tutti gli ordini temporali delle coppie delle alterazioni in P sono soddisfatte da T.
Più precisamente tre condizioni devono essere soddisfatte :

1. Per ogni coppia di alterazioni *a,b* in P, tale che *a* è un antenato di *b* in P, allora *a* è un antenato di *b* in T
2. Per tutte le coppie di alterazioni *a,b* in P contenute nello stesso nodo di P, queste appartengono allo stesso nodo in T
3. Per tutte le coppie di alterazioni *a,b* in P presenti in diverse diramazioni di P, queste appartengono a diverse diramazioni in T

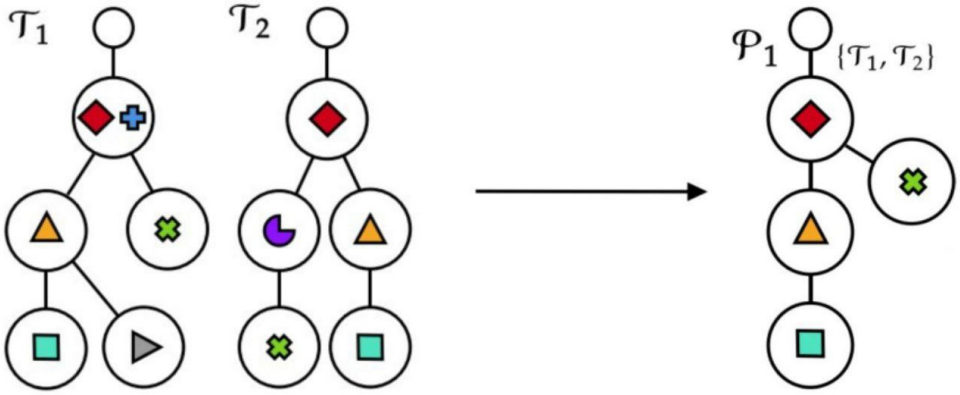


Figura 2.5: P_1 esempio di trajectory per i tumor tree T_1 e T_2 [1]

Da questo momento in avanti il fatto che una trajectory P è osservata in un tumor tree T verrà rappresentato come P ∈ T.
Un'ulteriore aspetto è che MASTRO non assume che le alterazioni in una trajectory siano consecutive nel tumore in cui queste sono osservate. Ad esempio nella Figura 2.5, si osserva che il diamante rosso e la croce verde nella trajectory P_1 sono consecutive per T_1, ma non per T_2.

Definizione 2.4 (Supporto di P)**.** Il supporto s_P di una trajectory P in un dataset $D = \{T_1, ..., T_n\}$ è il numero di alberi in cui P è osservata.

$$s_P = \sum_{i=1}^{n} \mathbb{1}[P \in T_i],$$

dove $\mathbb{1}[\cdot] = 1$ se l'argomento è vero, $\mathbb{1}[\cdot] = 0$ altrimenti.

Definizione 2.5 (Maximal trajectory)**.** Una trajectory P è **massimale** se aggiungendo una qualsiasi alterazione, non presente in P, il suo supporto s_P decresce.

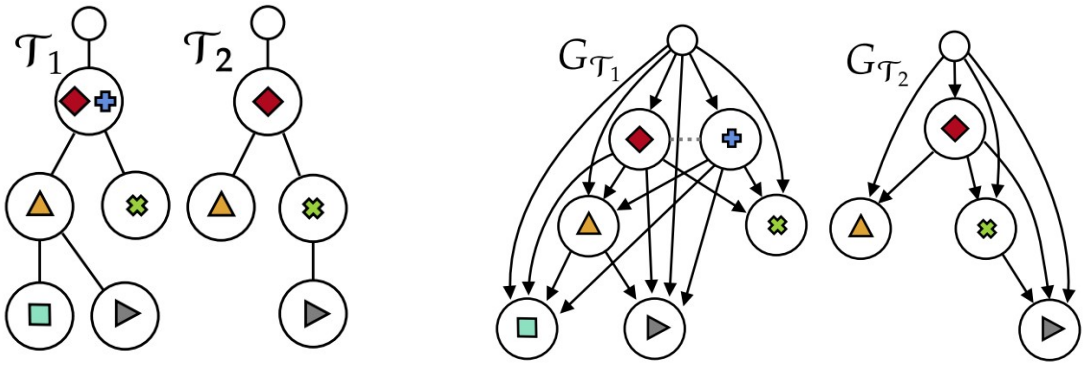
2.4 Rappresentazione tramite grafi

2.4.1 Expanded Tumor Graph

Un tumor tree T può essere rappresentato come un *expanded tumor graph*, definito come segue.

Definizione 2.6. Un expanded tumor graph $G_T = (V_T^G, E_T^G)$ di un tumor tree T è un grafo diretto con le seguenti proprietà:

- Per ogni alterazione $a \in$ A contenuta in un nodo $v \in V_T$, c'è un nodo $v_a \in V_T^G$ che contiene solo l'alterazione a.
- Per ogni coppia di alterazioni a,b di A, esiste un arco diretto $(v_a, v_b) \in E_T^G$ se e solo se a è un antenato di b (Definizione 2.2) in T.
- Per ogni coppia di alterazioni a,b che appartengono allo stesso nodo in T, esiste un arco chiamato **anti-edge** $(v_a, v_b, \star) \in E_T^G$, dove l'arco $\star$ indica che l'ordine tra a e b è sconosciuto.
- V_T^G contine un nodo vuoto v_r, il nodo radice di T, ed E_T^G contiene un arco diretto fra v_r e tutti gli altri nodi di G_T.



(a) tumor trees **(b)** expanded tumor graphs

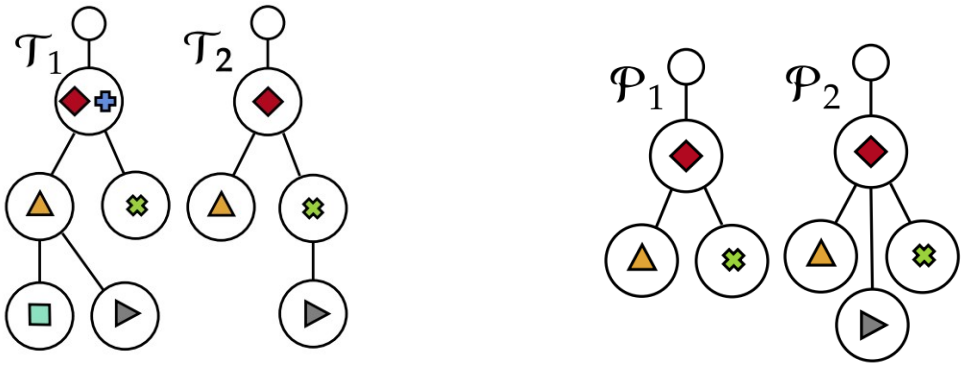
Figura 2.6: Expanded tumor graph G_{T_i} per il tumor tree T_i [1]

Tramite questa rappresentazione si può dedurre il seguente teorema:

Teorema 2.1. *Una trajectory P è osservata in T se e solo se G_P è un sottografo indotto di G_T.*

In questo modo, grazie al fatto che una trajectory P deve essere un sottografo indotto di un tumor tree T comporta che tutti gli ordini tra coppie di alterazioni sono preservate. Senza questo teorema infatti potrebbe accadere che alcune trajectory abbiano un ordine delle alterazioni parziale.

Ad esempio nella Figura 2.7 senza il requisito del sottografo indotto la trajecory P_2 è osservata in entrambi i tumor tree T_1 e T_2, in quanto l'ordine parziale delle alterazioni



(a) tumor trees **(b)** trajectories

Figura 2.7: P_2 esempio di trajectory errata [1]

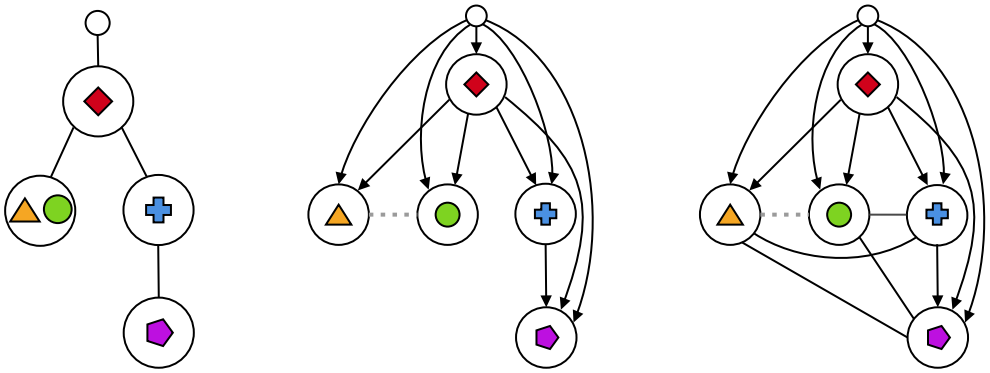
è soddisfatto. Tuttavia, la trajectory P_2 descriverebbe la mutazione grigia come una alterazione che non viene preceduta ne dalla alterazione arancione ne dalla alterazione verde, il che non è supportato da nessuno dei due tumor tree T_1 e T_2.
Quindi il Teorema 2.1 svolge in ruolo chiave affinché MASTRO riesca ad individuare l'ordine temporale in cui le mutazioni appaiono all'interno dei tumori.

2.4.2 Complete Tumor Graph

Per un tumor tree T si può specificare un insieme di archi $E_T^{\prime G} = (v_a, v_b, /)$ dove il simbolo / indica che i due nodi v_a e v_b, e quindi anche le loro rispettive alterazioni, appartengano a rami (detti anche branch) differenti del tumor tree T; da notare come questi archi non appartengono all'insieme E_T^G dell'expanded tumor graph di T, infatti:

$$E_T^{\prime G} = \{(v_a, v_b, /) : a \neq b, (v_a, v_b) \notin E_T^G, (v_b, v_a) \notin E_T^G, (v_a, v_b, \star) \notin E_T^G\}$$

Definizione 2.7. Un complete tumor graph è un grafo $G_T^c = (V_T^{G^c}, E_T^{G^c})$, dove $E_T^{G^c} = E_T^G \cup E_T^{\prime G}$.



(a) Tumor tree T **(b)** Expanded tumor graph G_T **(c)** Complete tumor graph G_T^c

Figura 2.8: Tutte le rappresentazioni per un dato tumor tree T

Quindi un grafo completo G_T^c ha 3 tipi di archi: diretti, non diretti fra mutazioni di rami diversi (/) e non diretti fra alterazioni nello stesso nodo in cui l'ordine non è conosciuto ($\star$).

Capitolo 3

Finding Frequent Maximal Trajectories

Il primo intento di MASTRO è quello di scoprire le *frequent maximal trajectories* (**FMT**) da n tumor trees; l'identificazione di queste trajectories tuttavia risulta computazionalmente complicato. In questo capitolo vedremo la formalizzazione del problema, le sue problematiche e come vengono risolte da MASTRO.

3.1 Frequent Maximal Trajectories Problem

Il problema computazionale di trovare le FMT viene formalizzato nel modo seguente:

Definizione 3.1 (FMT problem)**.** Dato un multiinsieme $\{T_1, ..., T_n\}$ con $n \geq 1$ tumor trees e una soglia $\sigma \in [1, n]$, la soluzione al problema sono tutte le traiettorie massimali (*maximal trajectories*) osservate in almeno σ tumor trees.

Quindi, il parametro σ da la libertà all'utente di controllare il minimo numero di tumor trees in cui una maximal trajectory è osservata.
L'identificazione di queste trajectories tuttavia risulta computazionalmente complicato; il FTM problem, anche se ristretto al caso $\sigma = n$, è un problema NP-hard [1]. Per risolvere questa inconvenienza, MASTRO ha ridotto il problema ad una variante del *frequent itemset mining problem*, un problema ben studiato in data mining che vanta di numerosi algoritmi efficienti.

3.2 Frequent Itemset Mining Problem

Il frequent itemset mining (**FIM**) è uno dei problemi più famosi nell'ambito del data mining e ha come obbiettivo quello di trovare pattern e regolarità all'interno di enormi dataset.
Le origini di questo problema risalgono alla conferenza internazionale SIGMOD del 1993 [9] nella quale venne presentata una soluzione iniziale per scoprire quali erano gli oggetti

acquistati più frequentemente dai consumatori. Lo scopo principale era quello di trovare schemi ricorrenti nel comportamento d'acquisto del consumatore nei supermercati; in pratica si cercava di identificare un insieme di oggetti che sono abitualmente comprati assieme.
Il FIM nel corso degli anni ha ampliato i suoi orizzonti e oggi le applicazioni sono numerose, dalla biomedica per trovare annotazioni (geni) che si presentano assieme, dal marketing per capire quali sono le abitudini dei consumatori e infine anche nell'ambito del text mining per scoprire frasi ricorrenti e associazioni di parole comuni. Questa estensione del problema in diversi ambiti è stata possibile grazie ad algoritmi che nel corso degli anni (Apriori, FP-Growth) hanno notevolmente migliorato l'efficienza e la scalabilità nel risolvere questo task.

Prima di dare una definizione formale del FIM problem è necessario descrivere alcuni concetti. Sia I un insieme di tutti i possibili items e una transazione t un sottoinsieme di I. Un dataset X è un multiinsieme di n transazioni $X = \{t_1, ..., t_n\}$. Un itemset A è un sottoinsieme di I e il suo insieme di supporto S(A) è un insieme di transazioni contenenti A: $S(A) = \{t_i, A \subseteq t_i, i \in [1, n]\}$. Il supporto s_A è definito come la cardinalità di S(A): $s_A = |S(A)|$. Il FIM problem è formalmente definito nel modo seguente:

Definizione 3.2 (FI problem)**.** Il problema del frequent itemset mining consiste nel calcolare l'insieme di itemsets con support $\geq \sigma$

$$FI(I, X, \sigma) = \Big\{ A \subseteq I : s_A \geq \sigma \Big\}$$

Symbol	Meaning
I	insieme di items
t	transazione
X	insieme di n transazioni
A	itemset
S(A)	insieme di supporto
s_A	supporto di A
σ	threshold
$FI(I, X, \sigma)$	frequent itemset mining problem

Tabella 3.1: Simbologia chiave per FIM

3.3 Frequent Trajectories

Per identificare le frequent maximal trajectories per un dataset $D = \{T_1, ..., T_n\}$ bisogna prima scoprire le traiettorie più frequenti da l'insieme degli archi più frequenti; questo obbiettivo può essere raggiunto applicando la funzione $FI(I, X, \sigma)$, per I e X appropriati, all'insieme di n rooted tumor trees $D = \{T_1, ..., T_n\}$.
I simboli che sono stati specificati nel nella Tabella 3.1 possono essere ridefiniti in linea con il problema che MASTRO risolve.
Definiamo I come l'unione tra gli insiemi di archi di tutti i complete tumor graphs (vedi def 2.7): $I = \cup_{T \in D} E_T^{G^c}$. Gli archi di ogni complete tumor graph di D sono un sottoinsieme di I, perciò definiamo $X = \{E_T^{G^c} : T \in D\}$, dove la i-esima transazione t_i equivale agli archi del i-esimo complete tumor graph $t_i = E_{T_i}^{G^c}$. Per un itemset $A \subseteq I$, $|A|$ rappresenta il numero di archi in A. L'insieme di nodi V(A) identifica i nodi che sono adiacenti ad almeno un arco di A:

$$V(A) = \Big\{ v : [\exists (v, w) \in A] \vee [\exists (w, v) \in A] \vee [\exists (w, v, l) \in A, l \in \{/, \star\}] \Big\}$$

Definizione 3.3 (FT)**.** L'insieme delle frequent trajectories $\mathbf{FT}(D, \sigma)$ è l'insieme di itemsets frequenti in X tale che $|A| = \binom{|V(A)|}{2}$.

$$FT(D, \sigma) = \left\{ A \in FI(I, X, \sigma) : |A| = \binom{|V(A)|}{2} \right\}$$

Il motivo per il quale $|A| = \binom{|V(A)|}{2}$ è che un itemset $A \subseteq I$, osservato in almeno $\sigma \geq 1$ complete tumor graphs, rappresenta l'insieme di archi di un sottografo completo con insieme di nodi V(A) se e solo se $|A| = \binom{|V(A)|}{2}$; questo deriva anche dal fatto che è un sottografo di almeno un tumor graph con archi, diretti o indiretti, tra coppie di alterazioni uniche appartenenti a V(A).

In accordo con la definizione di observed trajectory P (Definizione 2.3), le condizioni $A \subseteq E_T^{G^c}$ e $A \in FT(D, \sigma)$ implicano che l'insieme di archi A formi un sottografo indotto G_T, e viceversa. In altre parole, l'insieme di archi A è un sottoinsieme degli archi di un grafo completo ($A \subseteq E_T^{G^c}$) se e solo se tale insieme di archi forma un sottografo indotto G_T, ovvero tutti gli archi della traiettoria sono coerenti con gli archi del grafo del paziente, e non solamente una parte di questi. Perciò, esiste un'associazione unica tra un itemset $\in FT(D, \sigma)$ e una frequent trajectory; inoltre ogni itemset $\notin FT(D, \sigma)$ può essere tranquillamente scartato.

3.4 La soluzione di MASTRO

MASTRO risolve il problema di trovare le frequent maximal trajectories sfruttando la relazione che intercorre tra frequent itemsets e frequent trajectories.
In primis, MASTRO prende come input un multiinsieme $D = \{T_1, ..., T_n\}$ di tumor trees e una soglia minima $\sigma \in [1, n]$, e costruisce l'insieme X rispetto all'insieme I. Poi sfrutta un algoritmo per il frequent itemset mining per estrarre tutti gli itemsets frequenti $FI(I, X, \sigma)$. L'algoritmo prosegue nello scartare tutti gli itemsets A che rispettano la condizione $|A| \neq \binom{k}{2}$ per ogni valore di k, ottenendo così l'insieme $FT(D, \sigma)$. Questo perchè gli archi contenuti nell'itemset A devono rappresentare un grafo completo, quindi con un arco tra ogni coppia di alterazioni.
L'ultimo passaggio di MASTRO consiste nel calcolare l'insieme delle frequent maximal trajectories MFT(D, σ) da $FT(D, \sigma)$.

Teorema 3.1. *Per ogni elemento A di $FT(D, \sigma)$, se non c'è nessun altra frequent trajectory $A' \in FT(D, \sigma)$ con lo stesso insieme di supporto S(A) = S(A') e che $A' \supseteq A$, allora A è un complete tumor graph di una frequent maximal trajectory.*

Se questo teorema non venisse rispettato si potrebbero aggiungiere nodi e archi senza ridurre il supporto s_A di A, quindi in contrasto con la Definizione 2.5 di traiettoria massimale.
L'ultimo passaggio di MASTRO consiste nello stimare l'importanza delle traiettorie in $MFT(D, \sigma)$, descritto nel capitolo 4.

Capitolo 4

Significatività Statistica delle Traiettorie

Il secondo scopo di MASTRO è quello di misurare la significatività del supporto di una trajectory P. In questo capitolo verranno presentati alcuni concetti di probabilità ed in che modo l'algoritmo MASTRO riesce a raggiungiere questo traguardo.

4.1 Identificazione delle traiettorie significative

Per valutare l'importanza di un supporto di un trajectory P (Definizione 2.4) si considera quanto è probabile osservare P in un tumor tree T sotto l'ipotesi che le alterazioni in T siano casualmente assegnate ai suoi nodi. Dunque, MASTRO per valutare quanto un ordine preciso di alterazioni, essenzialmente una trajectory, sia probabile in un paziente, esegue un test statistico che condiziona su l'insieme di alterazioni di ciascun paziente. Per ottenere questo MASTRO calcola il valore atteso di alberi in cui una trajectory P è osservata e quanto è probabile osservare per caso una frequent trajectory.
L'importanza di una trajectory P viene calcolata considerando la probabilità di osservare P in ogni albero che viene generato da tre distribuzioni di probabilità in cui l'ipotesi nulla è vera. Per ogni albero T si considerino i seguenti tre modelli nulli:

1. **Independent assignment model**. Ogni alterazione di T viene assegnata a un nodo di T in modo indipendente e uniforme.

2. **Permutation assignment**. Le alterazioni di T vengono casualmente permutate tra i nodi di T; il numero di alterazioni per ogni nodi viene preservato.

3. **Independent assignment in random topology**. Dall'insieme dei tumor trees D viene casualmente scelta una topologia ed a questa vengono assegnate le alterazione di T in modo indipendente e uniforme.

Ogni modello nullo permette di calcolare la probabilità di osservare una trajectory P in un albero generato dal modello, $Pr(P \in T)$; per semplicità da adesso in avanti verranno

considerate le probabilità generate da un singolo modello nullo. Per vedere come utilizzare queste probabilità per stimare l'importanza di una trajectory, vi è la necessità di introdurre nuovi parametri.
Sia p_i la probabilità che una trajectory P sia osservata nel i-esimo tumor tree T, con $p_i = 0$ se le alterazioni A_P contenute nei nodi di P non sono un sottoinsieme di A_T.
Siano $X_1, ..., X_n$ delle variabili di Bernulli indipendenti tali che $Pr(X_i = 1) = E[X_i] = p_i$ e $Pr(X_i = 0) = 1 - p_i$. Definiamo X come una variabile binomiale di Poisson, dove X è la somma di tutti gli X_i: $X = \sum_{i=1}^{n} X_i$. La distribuzione binomiale di Poisson è il caso generico della distribuzione binomiale, in quanto le probabilità di successo p_i sono tutte diverse. Il valore atteso $E[X] = \sum_{i=1}^{n} p_i$ di X è il supporto atteso di una trajectory P sotto l'ipotesi nulla. Sia $I_P \subseteq [1, n]$ l'insieme di indici tale che $i \in I_P$ se e solo se $p_i > 0$.
Dunque, la probabilità $Pr(X \geq s_P)$ di osservare una trajectory P con supporto maggiore uguale a s_P, sotto l'ipotesi nulla, è pari a:

$$Pr(X \geq s_P) = \sum_{k=s_P}^{|I_P|} \sum_{J \subseteq I_P, |J|=k} \prod_{i \in J} p_i \prod_{j \notin J} (1 - p_i)$$

4.2 Rimozione dei falsi positivi

Una volta che le probabilità per una trajectory P sono state calcolate, MASTRO usa dei metodi di ricampionamento (resampling methods) per controllare la presenza di falsi positivi. Per controllare il Family-Wise Error Rate (**FWER**), ovvero la probabilità di riportare in output uno o più falsi positivi, MASTRO usa come test la Westfall and Young permutation procedure [10]. L'ultimo importante parametro che viene calcolato tramite una procedura basata sul ricampionamento [11] è il False Discovery Rate (**FDR**) [12], ovvero la percentuale attesa di false scoperte che sono riportate come significative.

Capitolo 5

Risultati Sperimentali

Questo capitolo presenta i risultati ottenuti dal'applicazione dell'algoritmo MASTRO su pazienti affetti da cancro al seno osservati in almeno $\sigma = 2$ tumor trees.

5.1 Dataset

Il dataset di riferimento proviene da uno studio clinico [2], nel quale i dati sono stati estratti tramite il sequenziamento *gene panel* su pazienti affetti da cancro al seno (breast cancer, **BC**). In seguito, il dataset è stato analizzato da due docenti dell'università dell'Illinois e reso disponibile nella loro pagina Github [3]; in totale sono presenti 1315 pazienti con 37809 alberi filogenetici. Tramite uno script python è stato calcolato il numero medio di alberi per paziente e lo scarto quadratico medio del numero di alberi per paziente, i cui valori sono rispettivamente 28.75 e 291.75; questi due risultati ci mostrano come la distribuzione di alberi sia eterogenea, si passa infatti da pazienti con un solo albero a casi dove un paziente può avere un massimo di 6332 alberi.
Vista la enorme quantità di alberi presenti, per diminuirne il numero si è deciso di scegliere casualmente, tramite un seed fissato, un solo albero filogenetico per ciascun paziente, dunque da 37809 alberi si è passati a 1315.

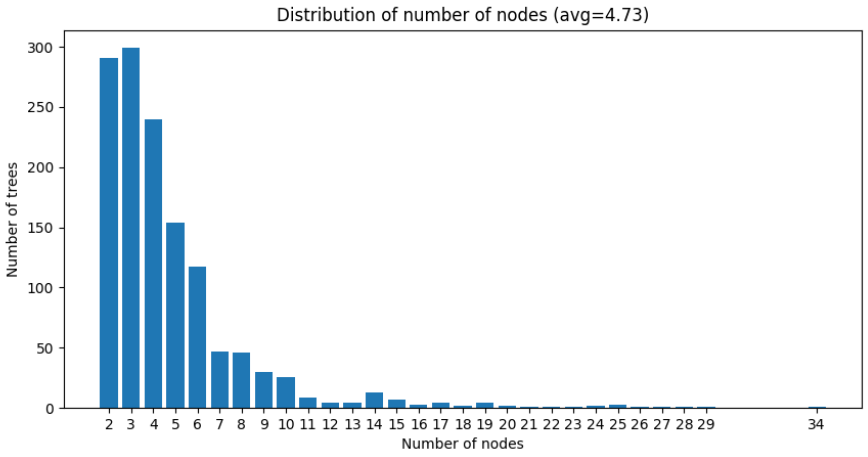


Figura 5.1: Distribuzione del numero di nodi per il BC dataset

In Figura 5.1 viene mostrata la distribuzione del numero di nodi, ovvero di mutazioni, per il dataset ristretto. In media nei 1315 alberi filogenetici selezionati sono presenti 4.73 alterazioni genetiche per albero.
Viste le caratteristiche di input per MASTRO, il passaggio successivo ha previsto il cambio di formato del dataset, infatti da alberi filogenetici si è passati alla loro rappresentazione tramite grafi completi (Definizione 2.7), sempre tramite uno script python; in Figura 5.2 è presente un esempio di tale conversione.

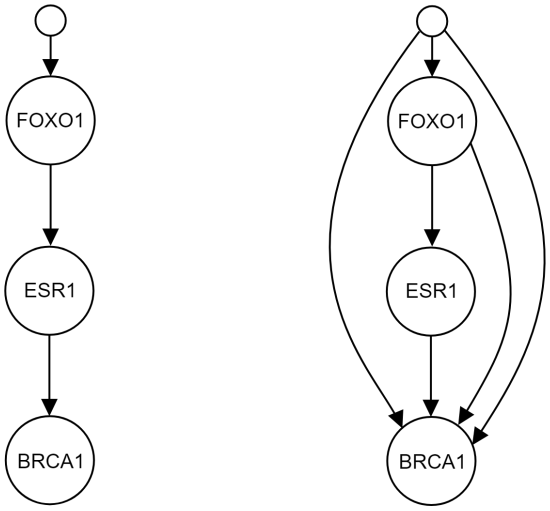


Figura 5.2: Esempio conversione di un albero del dataset considerato

5.2 Risultati

I test sono stati eseguiti su una macchina con processore Intel Xeon Gold 5220 2.2 GHz, 1 TB di RAM e con Ubuntu 20.04 come sistema operativo. I comandi eseguiti per ottenere i risultati presentati sono i seguenti:

1. `python3 run_MASTRO.py -g ../data/trees-bc.txt`
2. `python3 run_wy _correction.py -g ../data/trees-bc.txt -o PERM _bc_-m 1000 -minp BC_min.csv`
3. `python3 plot_results.py -r trees-bc_final.txt -minp BC_min.csv -t BC`
4. `python3 plot_emp_FDR.py -nullp PERM_bc_-r trees-bc_final.txt -t BC`

Il primo, terzo e quarto comando hanno tempi di esecuzione relativamente veloci. Il secondo comando corregge i falsi positivi utilizzando 10^3 permutazioni del BC dataset in circa 3 ore.
MASTRO trova 2567 maximal trajectories con almeno 2 mutazioni e osservate in almeno $\sigma = 2$ tumor trees. La Figura 5.3 mostra alcune importanti statistiche sulle 2567 maximal

trajectories trovate da MASTRO. In alto a sinistra vengono mostrate le probabilità (**p-values**) ordinate e le 3 soglie derivanti dalla correzione di WY utilizzando come modello di permutazione quello di default, ovvero indipendent, per $m = 10^3$ permutazioni come paramentro; i valori delle soglie sono:

- $4.4 \cdot 10^{-}8$ per $\alpha = 0.01$
- $2.2 \cdot 10^{-}7$ per $\alpha = 0.05$
- $7.0 \cdot 10^{-}7$ per $\alpha = 0.1$

Dunque, si osserva che 3 traiettorie sono significative con FWER ≤ 0.01, 5 con FWER$\leq$ 0.05 e 9 con FWER$\leq$ 0.1.
Tra le altre figure presenti viene mostrato il supporto delle traiettorie rispetto al numero di alterazioni (Figura 5.3 alto a destra), i p-values delle traiettorie contro il numero di alterazioni (Figura 5.3 basso a sinistra), ed i p-values delle traiettorie versus il supporto di queste (Figura 5.3 basso a destra).

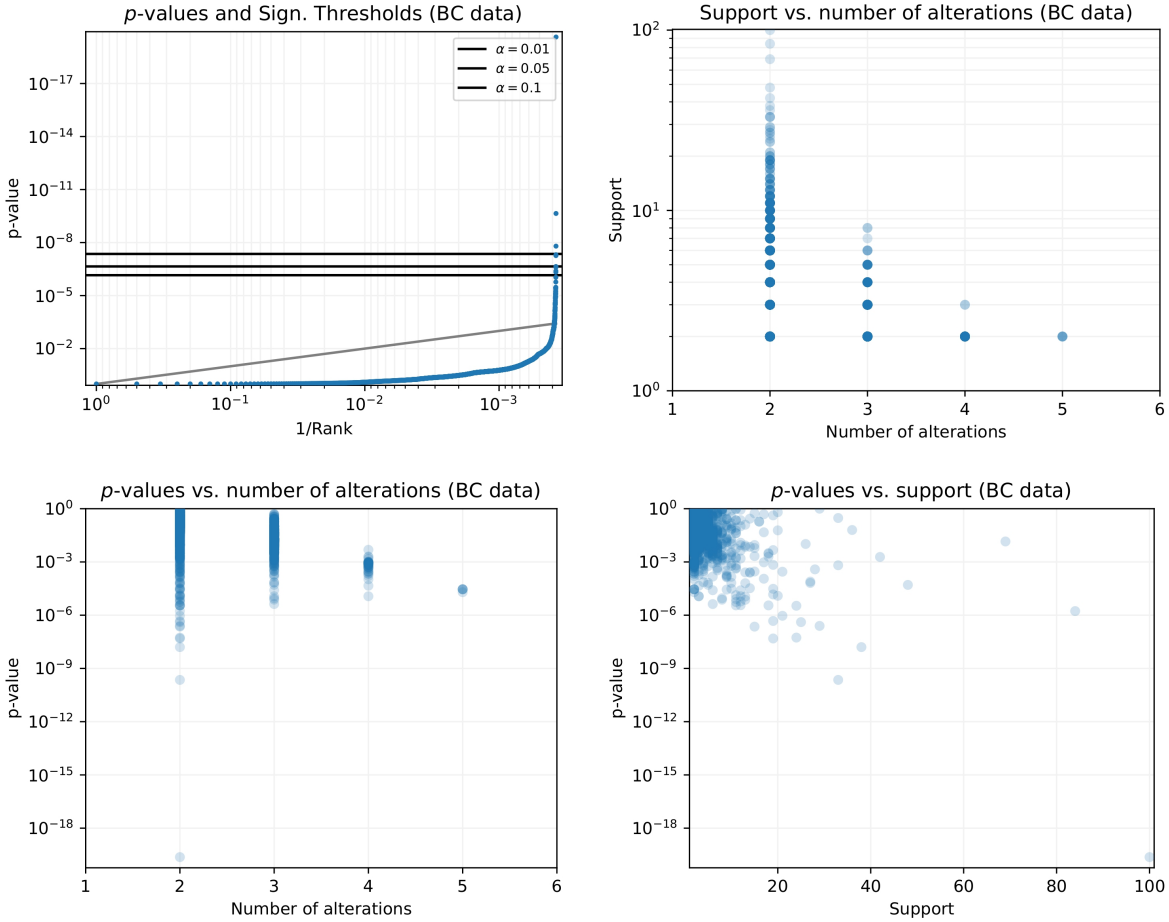


Figura 5.3: Statistiche dal BC dataset

MASTRO oltre a trovare le traiettorie più significative riesce anche a stimare il valore del FDR dei primi k risultati migliori ordinati per p-value crescenti. Più il valore del FDR è basso (ad esempio ≤ 0.2) e meno sono presenti false scoperte nelle traiettorie considerate;

il numero di trajectories per cui il $FDR = 0.2$ ($\frac{1}{5}$ delle traiettorie sono potenzialmente falsi positivi) è pari a 76. Nella Figura 5.4 viene messo in luce come più traiettorie si considerano più la percentuale di errore aumenta.

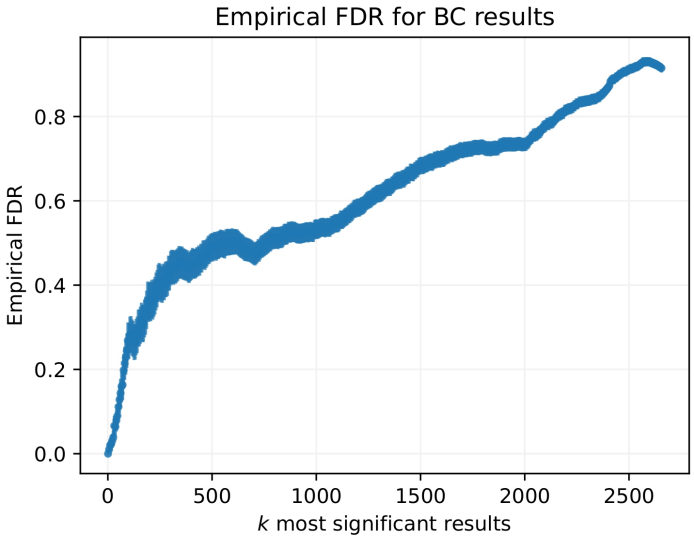
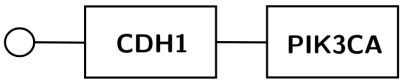


Figura 5.4: Stima empirica del FDR per i k risultati più significativi

L'analisi dei risultati si concentra sulle prime 10 maximal trajectories più significative. In Figura 5.5 sono presenti le traiettorie in questione, i p-values, il supporto ed i tumor trees in cui sono osservate.

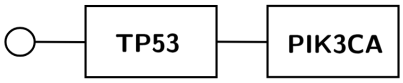
1) $p = 2.34 \cdot 10^{-20}$ $s = 100$

{3, 12, 25, 35, 39, 48, 50, 61, 117, 127, 136, 151, 175, 208, 211, 212, 215, 217, 227, 243, 249, 254, 277, 300, 308, 322, 331, 339, 347, 352, 374, 419, 421, 426, 427, 455, 466, 503, 507, 511, 528, 530, 534, 549, 561, 562, 564, 565, 592, 601, 602, 620, 656, 659, 677, 711, 712, 718, 730, 742, 747, 784, 823, 829, 860, 882, 898, 901, 912, 928, 930, 932, 937, 972, 991, 1007, 1017, 1024, 1055, 1080, 1082, 1090, 1108, 1110, 1112, 1128, 1133, 1149, 1165, 1203, 1224, 1237, 1250, 1266, 1276, 1283, 1285, 1287, 1290, 1308}



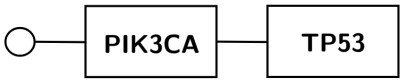
2) $p = 2.26 \cdot 10^{-10}$ $s = 84$

{22, 25, 41, 62, 68, 83, 100, 106, 136, 166, 169, 238, 258, 262, 267, 281, 282, 292, 293, 314, 327, 388, 425, 426, 438, 473, 503, 506, 521, 525, 533, 536, 542, 544, 566, 570, 581, 621, 634, 656, 670, 681, 682, 714, 723, 725, 731, 743, 750, 771, 784, 789, 790, 796, 804, 812, 813, 829, 847, 869, 894, 902, 906, 951, 1005, 1012, 1019, 1022, 1033, 1036, 1037, 1051, 1078, 1118, 1137, 1151, 1157, 1166, 1169, 1207, 1238, 1244, 1286, 1296}



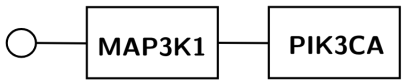
3) $p = 1.59 \cdot 10^{-8}$ $s = 69$

{57, 70, 76, 80, 90, 107, 114, 116, 148, 158, 190, 209, 215, 260, 269, 285, 296, 298, 304, 312, 330, 351, 360, 380, 389, 401, 493, 509, 543, 547, 572, 584, 585, 601, 604, 606, 635, 637, 642, 646, 655, 664, 677, 705, 744, 770, 800, 853, 892, 942, 946, 948, 980, 990, 1016, 1059, 1064, 1093, 1114, 1140, 1170, 1193, 1200, 1229, 1243, 1259, 1298, 1304, 1314}



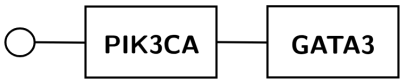
4) $p = 4.86 \cdot 10^{-8}$ $s = 48$

{7, 8, 22, 48, 62, 72, 98, 121, 170, 177, 187, 217, 218, 248, 307, 321, 365, 378, 386, 426, 479, 613, 646, 666, 676, 712, 751, 758, 772, 792, 808, 817, 886, 888, 950, 958, 1033, 1056, 1068, 1074, 1079, 1099, 1107, 1135, 1144, 1148, 1287, 1297}



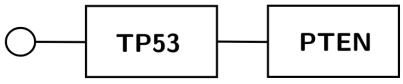
5) $p = 5.47 \cdot 10^{-8}$ $s = 42$

{38, 89, 91, 114, 154, 176, 239, 266, 270, 277, 281, 286, 313, 375, 376, 498, 532, 549, 572, 635, 675, 677, 679, 724, 754, 761, 780, 786, 835, 953, 994, 1001, 1005, 1014, 1022, 1038, 1111, 1148, 1200, 1204, 1270, 1307}



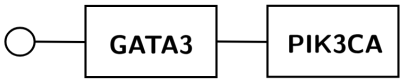
6) $p = 2.27 \cdot 10^{-7}$ $s = 38$

{113, 120, 125, 173, 314, 318, 389, 411, 415, 436, 458, 508, 513, 514, 520, 538, 542, 577, 603, 622, 628, 640, 642, 733, 763, 800, 816, 846, 925, 990, 993, 1012, 1015, 1051, 1064, 1194, 1200, 1234}



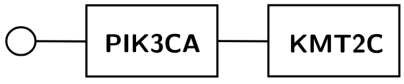
7) $p = 2.49 \cdot 10^{-7}$ $s = 36$

{37, 72, 78, 145, 153, 160, 170, 187, 229, 240, 245, 307, 363, 364, 378, 467, 511, 604, 611, 639, 717, 745, 781, 817, 896, 959, 971, 1000, 1021, 1042, 1079, 1122, 1135, 1169, 1186, 1277}



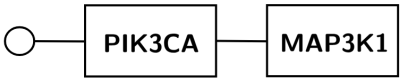
8) $p = 4.07 \cdot 10^{-7}$ $s = 33$

{25, 127, 168, 197, 296, 308, 352, 376, 379, 386, 463, 503, 533, 572, 608, 682, 709, 773, 853, 854, 862, 870, 934, 950, 960, 971, 1064, 1074, 1203, 1237, 1278, 1279, 1288}



9) $p = 4.73 \cdot 10^{-7}$ $s = 33$

{25, 35, 102, 134, 147, 179, 197, 266, 277, 331, 374, 380, 400, 470, 509, 576, 667, 745, 773, 819, 840, 856, 892, 934, 971, 972, 1042, 1110, 1165, 1186, 1238, 1278, 1301}



10) $p = 9.25 \cdot 10^{-7}$ $s = 33$

{153, 154, 248, 266, 281, 307, 314, 363, 549, 608, 613, 639, 652, 738, 742, 774, 781, 808, 912, 958, 991, 994, 1016, 1017, 1033, 1055, 1088, 1101, 1121, 1135, 1179, 1266, 1277}

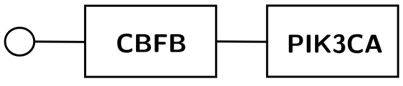


Figura 5.5: Le 10 traiettorie più significative trovate da MASTRO per il BC dataset

Come si può notare, la struttura di tutte le 10 migliori trajectory è molto semplice (formata da $Germline \rightarrow Mutation1 \rightarrow Mutation2$).

Un ulteriore fattore comune nei risultati è che le coppie di trajectory con rango (2,3), (4,9), (5,7) hanno le stesse mutazioni ma in ordine invertito; quest'ultima caratteristica ci indica come TP53 e PIK3CA, MAP3K1 e PIK3CA, PIK3CA e GATA3 siano alterazioni fortemente connesse e intercambiabili.

5.2.1 Confronto con CloMu

I risultati sono stati confrontati con l'analisi presentata nel paper delle fonti del dataset, in particolare nella sezione 4.2 di [3]. Rispetto ai risultati riportati da MASTRO, si possono sottolineare similitudini e differenze.
L'articolo [3] indica che le più forti coppie di mutazioni intercambiabili sono due: CDH1 e PIK3CA, GATA3 e MAP3K1. La prima di queste coppie corrisponde con la trajectory calcolata da MASTRO al primo posto (rank R=1), mentre la sua controparte non risulta essere nelle prime 10 posizioni (presente al rank R=14). La seconda coppia non risulta intercambiabile e nemmeno presente come trajectory secondo quanto calcolato da MASTRO entro le 50 trajectories più significative. In aggiunta, MASTRO calcola che MAP3K1 e PIK3CA siano mutazioni intercambiabili (traiettorie presenti ai rank 4 e 9), dove per [3] questa coppia non risulta essere intercambiabile.
Secondo [3] la mutazione TP53, una delle alterazioni più presenti nel dataset, sarebbe un caso particolare in quanto avrebbe bassi livelli di causalità tra le mutazioni più ricorrenti (CDH1, PIK3CA, GATA3, MAP3KI), mentre avrebbe alti livelli con altre mutazioni. MASTRO, invece, osserva una forte relazione tra le mutazioni TP53 e PIK3CA in quanto queste due mutazioni sarebbero la seconda e la terza traiettoria per importanza nel dataset.
Un ulteriore dato che è possibile confrontare riguarda le relazioni temporali tra 2 mutazioni, ovvero quanto una mutazione è causata da un'altra. Tra le relazioni presenti in [3] si cita che CDH1 e GATA3 causano PIK3CA; entrambe queste relazioni sono presenti in Figura 5.5 rispettivamente alle posizioni 1 e 7.
Molte sono le traiettorie calcolate da MASTRO che non risultano significative oppure presenti per [3]; queste traiettorie potrebbero fornire nuove traiettorie di alterazioni di interesse biologico.

5.3 Conclusioni e Future Works

In conclusione la traiettoria più significativa e coerente con altri articoli risulta essere **Germline** $\rightarrow$ **CDH1** $\rightarrow$ **PIK3CA**, ovvero la trajectory al primo posto secondo quanto calcolato da MASTRO.
Tra i possibili miglioramenti futuri vengono elencati:

1. Effettuare un'analisi più approfondita del dataset attuale tenendo in considerazione più trajectory, non limitandosi alle migliori 10. Infatti, MASTRO indica che le 76 traiettorie maggiormente significative hanno FDR non elevato (≤ 0.2).

2. Eseguire in parallelo una analisi su $m = 10^4$ permutazioni tramite la WY permutation procedure e confrontare i nuovi valori di FDR e FWER con i precedenti.

3. Riuscire ad integrare insiemi di alberi per ciascun paziente.

4. Analizzare e confrontare i risultati con diverse pubblicazioni da quelle considerate. In questa maniera sarebbe possibile avere una visione generale, al fine di validare le relazioni tra le mutazioni identificate.

5. Utilizzare altri dataset di pazienti affetti da cancro al seno per poter confrontare i nuovi risultati con quelli attuali.

Bibliografia

[1] Leonardo Pellegrina, Fabio Vandin (2022) *Discovering significant evolutionary trajectories in cancer phylogenies*, Bioinformatics.

[2] Razavi, P., Chang, M. T., Xu, G., Bandlamudi, C., Ross, D. S., Vasan, N., ... & Baselga, J. (2018). The genomic landscape of endocrine-resistant advanced breast cancers. Cancer cell, 34(3), 427-438.

[3] Ivanovic, S., El-Kebir, M. (2023). Modeling and predicting cancer clonal evolution with reinforcement learning. Genome Research, gr-277672.

[4] `https://www.airc.it` *Associazione italiana per la ricerca sul cancro*

[5] Tarabichi, M., Salcedo, A., Deshwar, A. G., Ni Leathlobhair, M., Wintersinger, J., Wedge, D. C., Boutros, P. C. (2021). A practical guide to cancer subclonal reconstruction from DNA sequencing. Nature methods, 18(2), 144-155.

[6] Turajlic, S., Sottoriva, A., Graham, T., Swanton, C. (2019). Resolving genetic heterogeneity in cancer. Nature Reviews Genetics, 20(7), 404-416.

[7] Apostoli, Anthony, Ailles, Laurie. (2015). Clonal evolution and tumor-initiating cells: New dimensions in cancer patient treatment. Critical reviews in clinical laboratory sciences. 53. 1-12. 10.3109/10408363.2015.1083944.

[8] Davis, A., Gao, R., Navin, N. (2017). Tumor evolution: Linear, branching, neutral or punctuated?. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1867(2), 151-161.

[9] Agrawal, R., Imieliński, T., Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (pp. 207-216).

[10] Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment (Vol. 279). John Wiley & Sons.

[11] Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), 9440-9445.

[12] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

[13] Luna, J. M., Fournier-Viger, P., Ventura, S. (2019). Frequent itemset mining: A 25 years review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(6), e1329.

[14] Giannini, S. (2016). Strumenti statistici per elaborazione dati su sequenziamenti di genoma umano (Images at page 8 and 9).

[15] Yikrazuul, (2008). "Base pair GC.svg", "Base pair AT.svg" [Images]. Retrieved from https://commons.wikimedia.org/wiki/File:Base_pair_GC.svg, https://commons.wikimedia.org/wiki/File:Base_pair_AT.svg

[16] Image retrived from https://www.dnaexpress.it/glossario/gene/