A Comparative Analysis of Algorithms for Identifying Cancer Driver Pathways

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica
Corso di Laurea in Informatica

SAPIENZA
UNIVERSITÀ DI ROMA

Candidato: Alessio Bandiera - 1985878

Relatore: Ivano Salvo

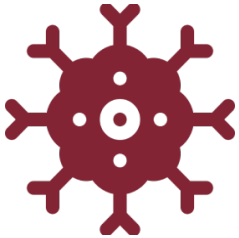
Anno Accademico: 2023/2024

Il cancro

Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.

Il cancro

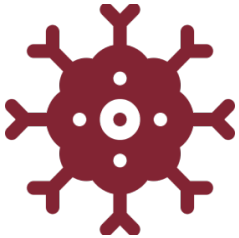
Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.



Esistono oltre 100 tipi di cancro, e.g. carcinomi, sarcomi e leucemie.

Il cancro

Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.

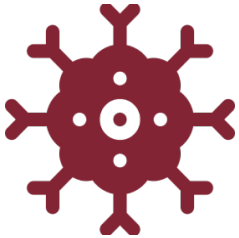


Esistono oltre 100 tipi di cancro, e.g. carcinomi, sarcomi e leucemie.

Ogni anno i decessi per il cancro sono nell'ordine dei milioni.

Il cancro

Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.



Esistono oltre 100 tipi di cancro, e.g. carcinomi, sarcomi e leucemie.

Ogni anno i decessi per il cancro sono nell'ordine dei milioni.

È fondamentale trovare trattamenti efficaci contro questa malattia.

Cure attuali

Le cure ed i trattamenti per il cancro attualmente disponibili sono:

Chirurgia

Radioterapia

Chemioterapia

Terapie ormonali

Cure attuali

Le cure ed i trattamenti per il cancro attualmente disponibili sono:

Chirurgia

Radioterapia

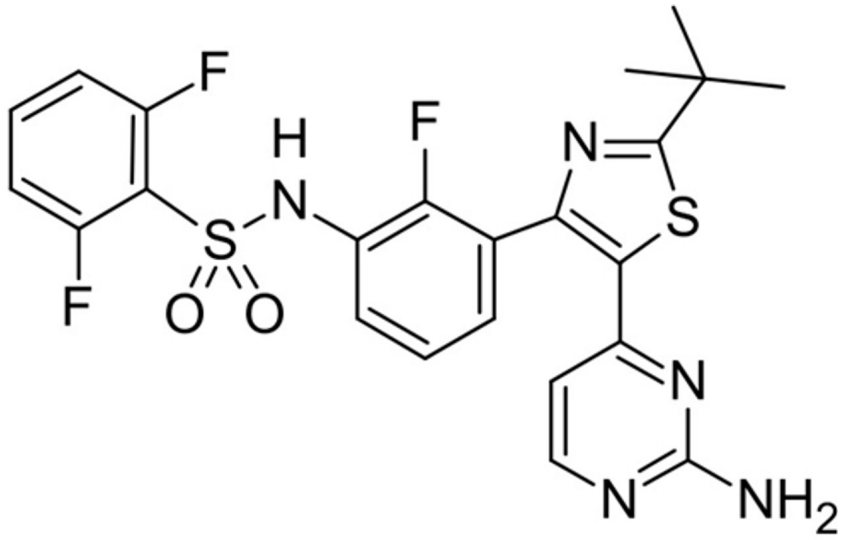
Chemioterapia

Terapie ormonali

Problema. Tutti i trattamenti attuali sono limitati e possono portare a molteplici effetti collaterali.

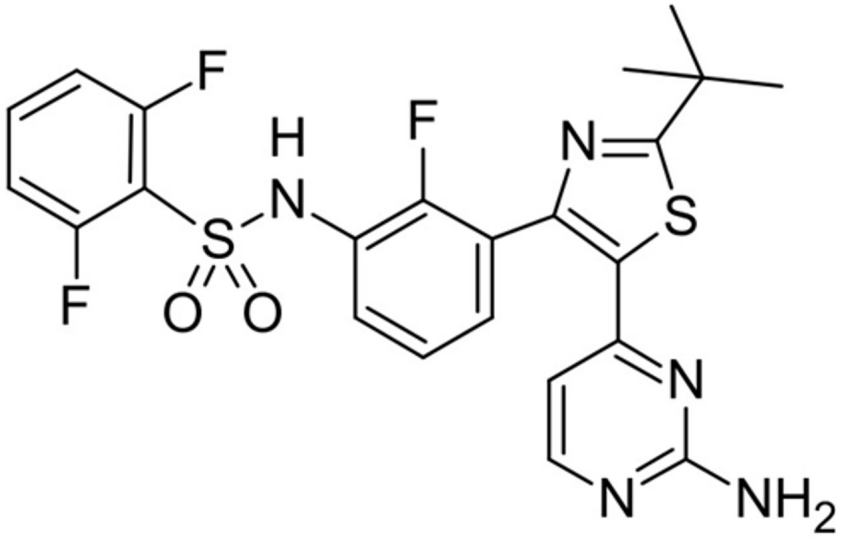
Terapia a bersaglio

La **terapia a bersaglio** è un trattamento per il cancro che si concentra sulle proteine responsabili della crescita del tumore.



Terapia a bersaglio

La **terapia a bersaglio** è un trattamento per il cancro che si concentra sulle proteine responsabili della crescita del tumore.

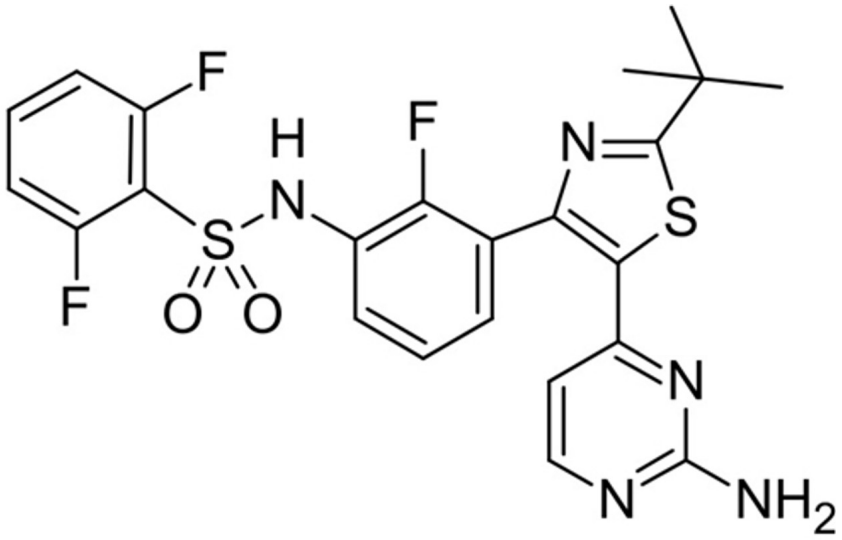


⇓

La terapia a bersaglio offre maggiore selettività e può aiutare a ridurre gli effetti collaterali.

Terapia a bersaglio

La **terapia a bersaglio** è un trattamento per il cancro che si concentra sulle proteine responsabili della crescita del tumore.



⇓

La terapia a bersaglio offre maggiore selettività e può aiutare a ridurre gli effetti collaterali.

⇓

Cosa bersagliare?

Il ruolo delle mutazioni nel cancro

Lo sviluppo del cancro è un **processo di mutazione** e selezione di cellule con capacità sempre maggiori di proliferare.

Il ruolo delle mutazioni nel cancro

Lo sviluppo del cancro è un **processo di mutazione** e selezione di cellule con capacità sempre maggiori di proliferare.

⇓

Le **mutazioni** ricoprono un ruolo fondamentale per lo sviluppo e la progressione del cancro.

Tipi di mutazioni

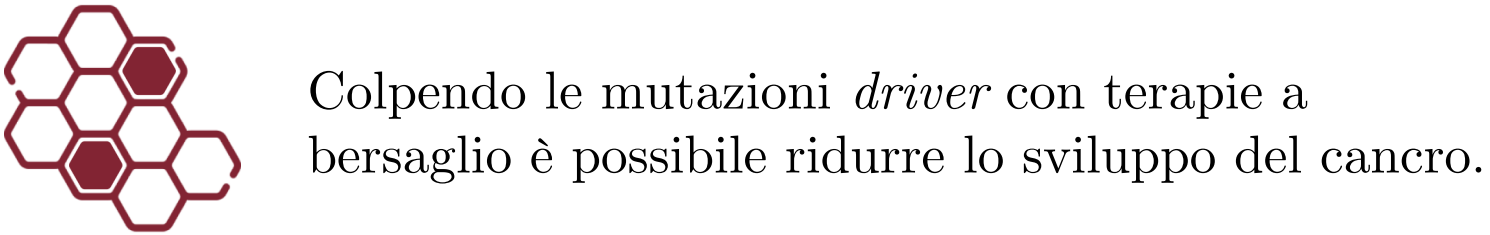
Una **mutazione *passenger*** è una mutazione che non conferisce vantaggio diretto allo sviluppo del cancro.

Una **mutazione *driver*** è una mutazione che contribuisce direttamente alla crescita tumorale.

Tipi di mutazioni

Una **mutazione *passenger*** è una mutazione che non conferisce vantaggio diretto allo sviluppo del cancro.

Una **mutazione *driver*** è una mutazione che contribuisce direttamente alla crescita tumorale.



Colpendo le mutazioni *driver* con terapie a bersaglio è possibile ridurre lo sviluppo del cancro.

Tipi di mutazioni

Una **mutazione *passenger*** è una mutazione che non conferisce vantaggio diretto allo sviluppo del cancro.

Una **mutazione *driver*** è una mutazione che contribuisce direttamente alla crescita tumorale.

⇓

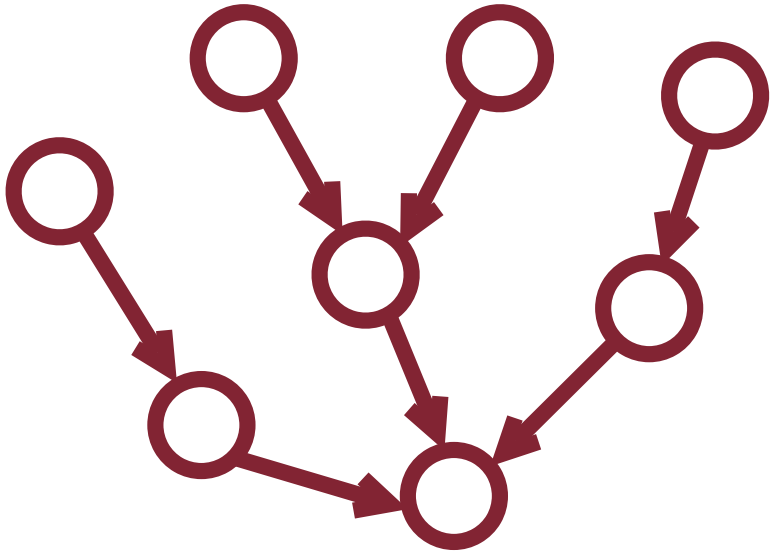
Colpendo le mutazioni *driver* con terapie a bersaglio è possibile ridurre lo sviluppo del cancro.

⇓

Classificare le mutazioni tra *driver* e *passenger* è essenziale.

Pathway cellulari

Un **pathway** cellulare è insieme di catene di processi biochimici che avvengono in una cellula.

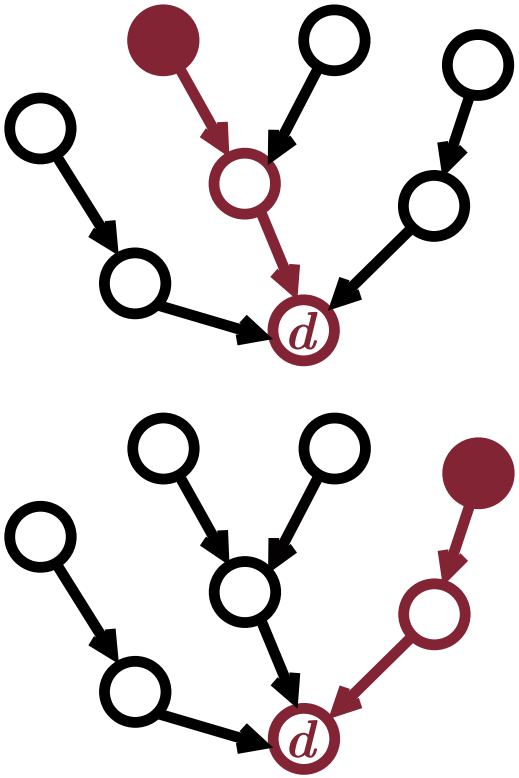


I pathway possono essere rappresentati da grafi diretti.

Siamo interessati ai geni che compongono i pathway.

Cercare i pathway *driver*

I pathway sono importanti poiché nel loro contesto è possibile classificare le mutazioni del cancro.



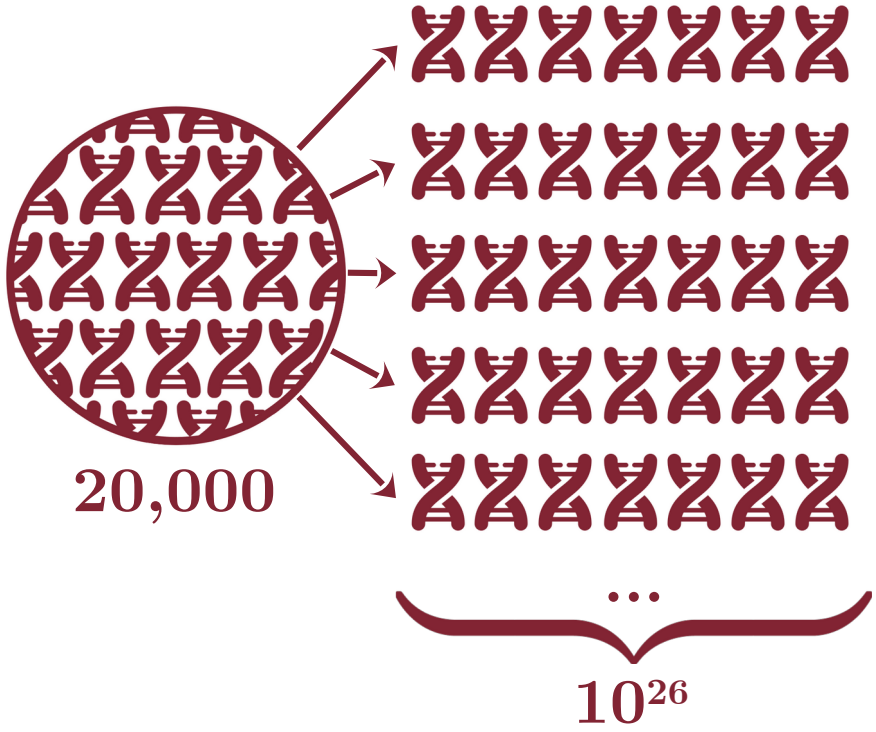
Più mutazioni *driver* in geni diversi possono portare a simili effetti *downstream*.

⇓

Mutazioni diverse possono influenzare lo stesso pathway in vari campioni.

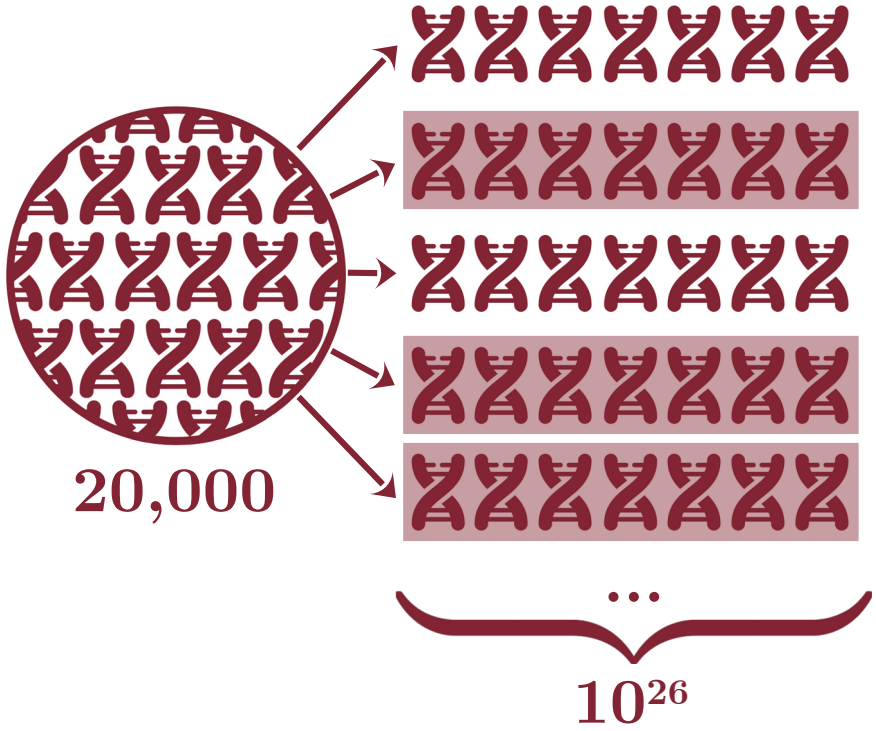
Problemi nel cercare i pathway

Problema. Cercare pathway *driver* è complesso, per via dell'enorme numero di pathway possibili da verificare.



Problemi nel cercare i pathway

Problema. Cercare pathway *driver* è complesso, per via dell'enorme numero di pathway possibili da verificare.



Non è possibile controllare ogni pathway.

Fortunatamente, statisticamente si osservano proprietà che permettono di ridurre il numero di pathway da controllare.

Formalizzazione dei pathway *driver*

Copertura: i geni *driver* di pathway *driver* sono mutati nella maggior parte dei pazienti.

Mutua esclusività: i geni *driver* all'interno dello stesso pathway sono approssimativamente mutuamente esclusivi.

Un **pathway *driver*** è costituito da geni mutati in numerosi pazienti, e le cui mutazioni sono approssimativamente mutuamente eslcusive all'interno del pathway.

Matrice di mutazione

Definizione. Una matrice di mutazione è una matrice binaria che descrive le mutazioni dei pazienti.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

Copertura di un gene

Definizione. La copertura di un gene g è l'insieme dei pazienti che hanno g mutato.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$\Gamma(g) := \{i \mid a_{i,j} = 1\}$$

Copertura di un insieme di geni

Definizione. La copertura di un insieme di geni M è l'unione delle coperture dei geni di M.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$\Gamma(M) := \bigcup_{g \in M} \Gamma(g)$$

Mutua esclusività

Definizione. M è mutuamente esclusivo se non ci sono pazienti con più di una mutazione di geni di M.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$\forall g, g' \in M \quad \Gamma(g) \cap \Gamma(g') = \varnothing$$

Sovrapposizione di un insieme di geni

Definizione. $\omega(M)$ rappresenta il numero di pazienti con più di un gene di M mutato.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$\omega(M) := \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$$

Sovrapposizione di un insieme di geni

Definizione. $\omega(M)$ rappresenta il numero di pazienti con più di un gene di M mutato.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$\omega(M) := \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$$

Sovrapposizione di un insieme di geni

Definizione. $\omega(M)$ rappresenta il numero di pazienti con più di un gene di M mutato.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$\omega(M) := \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$$

Sovrapposizione di un insieme di geni

Definizione. $\omega(M)$ rappresenta il numero di pazienti con più di un gene di M mutato.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$\omega(M) := \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$$

Peso di un gruppo di geni

Definizione. Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

Peso di un gruppo di geni

Definizione. Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

Peso di un gruppo di geni

Definizione. Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

Peso di un gruppo di geni

Definizione. Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

Maximum Weight Submatrix Problem (MWSP)

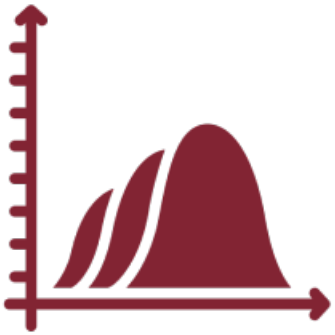
Definizione. (MWSP) Data una matrice di mutazione A di dimensioni $m \times n$, ed un intero $k > 0$, si trovi una sottomatrice $m \times k$ di A tale da massimizzare $W(M)$.

	g_1	g_2	g_3	g_4	g_5
p_1	0	1	0	0	1
p_2	1	0	1	0	0
p_3	0	1	1	0	0
p_4	0	0	0	1	1

Teorema. (MWSP) L'MWSP è NP-completo.

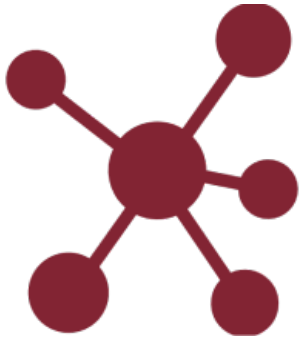
Approcci alla ricerca dei pathway

ILP



Statistica

Algoritmo genetico



Clustring

Approccio con ILP

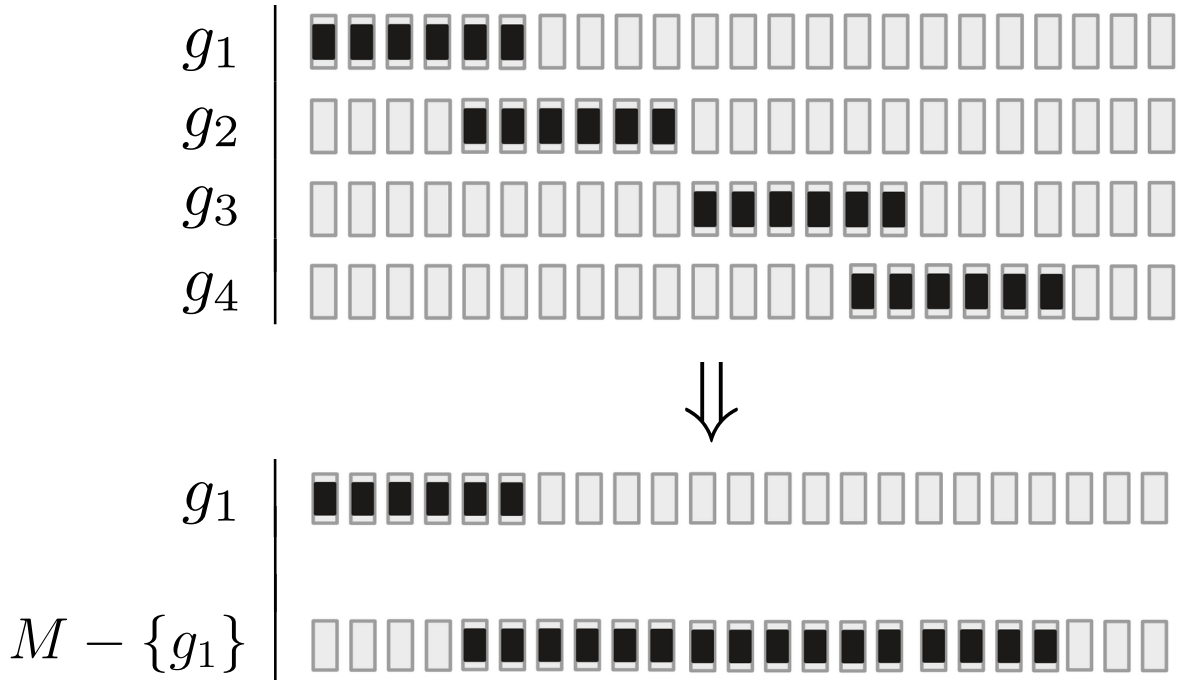
$$\text{maximize} \sum_{i=1}^{m} \left(2 \cdot C_i(M) - \sum_{j=1}^{n} I_M(j) \cdot a_{i,j} \right),$$

$$\text{subject to} \sum_{j=1}^{n} I_M(j) = k,$$

$$\sum_{j=1}^{n} I_M(j) \cdot a_{i,j} \geq C_i(M), \quad 1 \leq i \leq m.$$

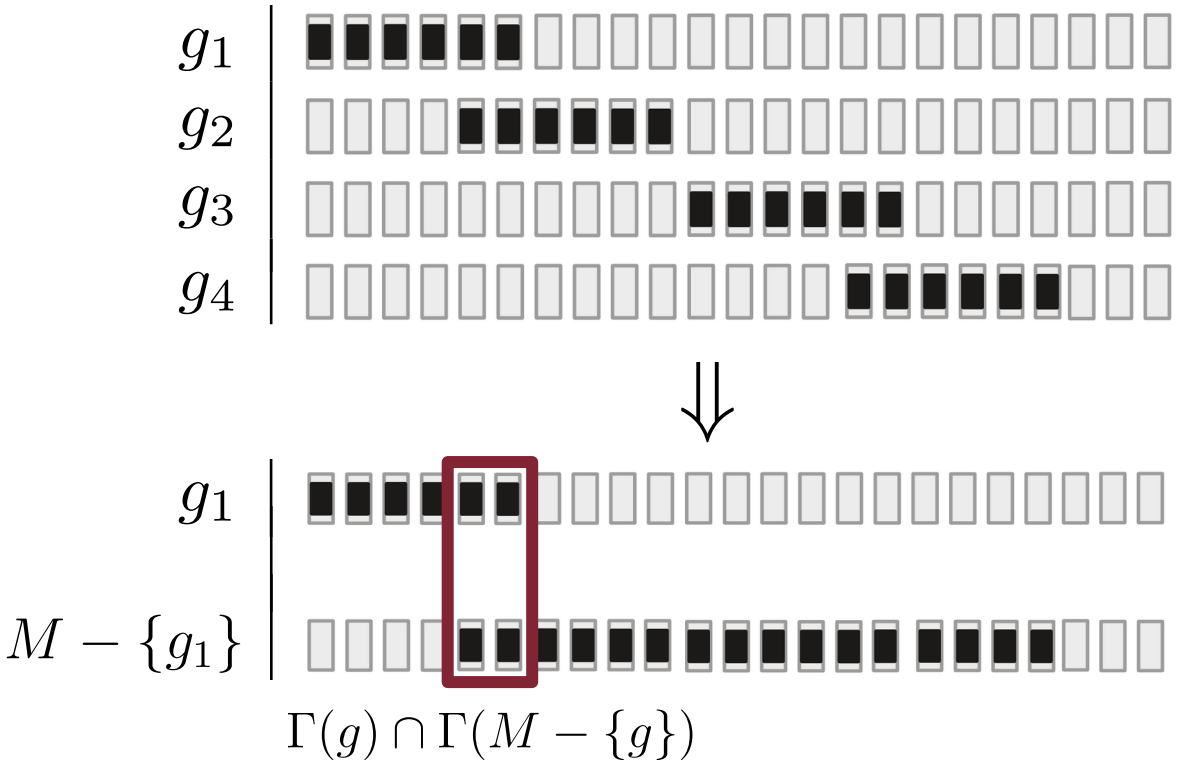
Approccio probabilistico

Ipotesi nulla: Dato un gruppo di geni M, un gene g di M è alterato indipendentemente dall'unione delle alterazioni dei geni in $M - \{g\}$.



Approccio probabilistico

Ipotesi nulla: Dato un gruppo di geni M, un gene g di M è alterato indipendentemente dall'unione delle alterazioni dei geni in $M - \{g\}$.



Approccio probabilistico

Ipotesi nulla: Dato un gruppo di geni M, un gene g di M è alterato indipendentemente dall'unione delle alterazioni dei geni in $M - \{g\}$.

$$X \sim H(m, \Gamma(g), \Gamma(M - \{g\}))$$

$$\Downarrow$$

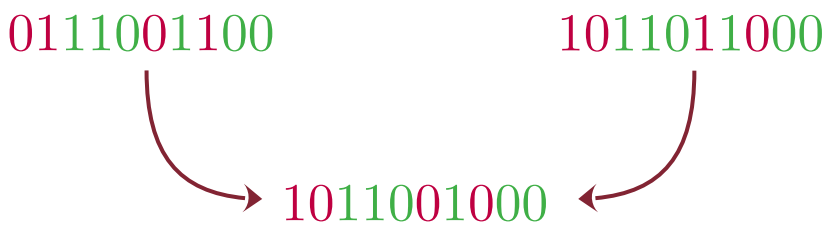
$$s_M := \max_{g \in M} P(X = \Gamma(g) \cap \Gamma(M - \{g\}))$$

Approccio genetico

L'algoritmo genetico utilizza la stessa funzione di *fitness* $W(M)$.

g_1	0111001100
g_2	1011011000
g_3	0001110011
g_4	0101010110
	⋮
g_n	0101010110
	⋮

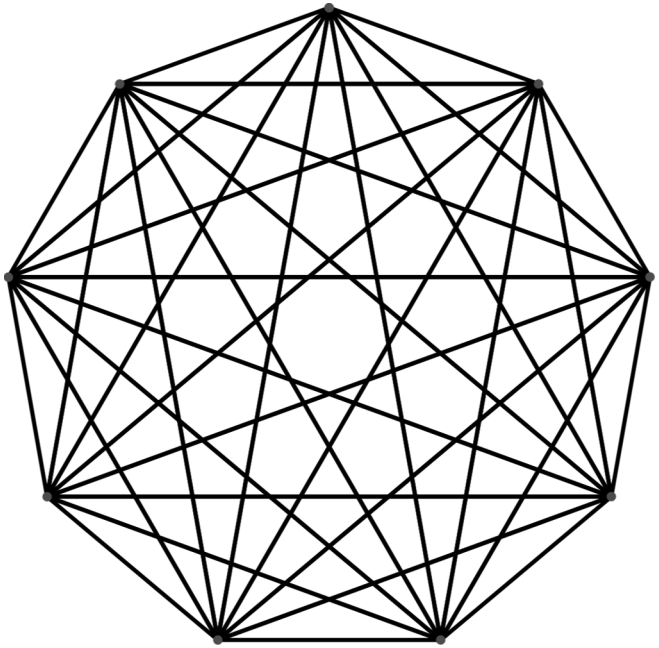
Un **membro** della popolazione è una stringa binaria che rappresenta un insieme di geni.



Un figlio eredita dai genitori i bit in comune, mentre gli altri sono casuali.

Approccio di clustering

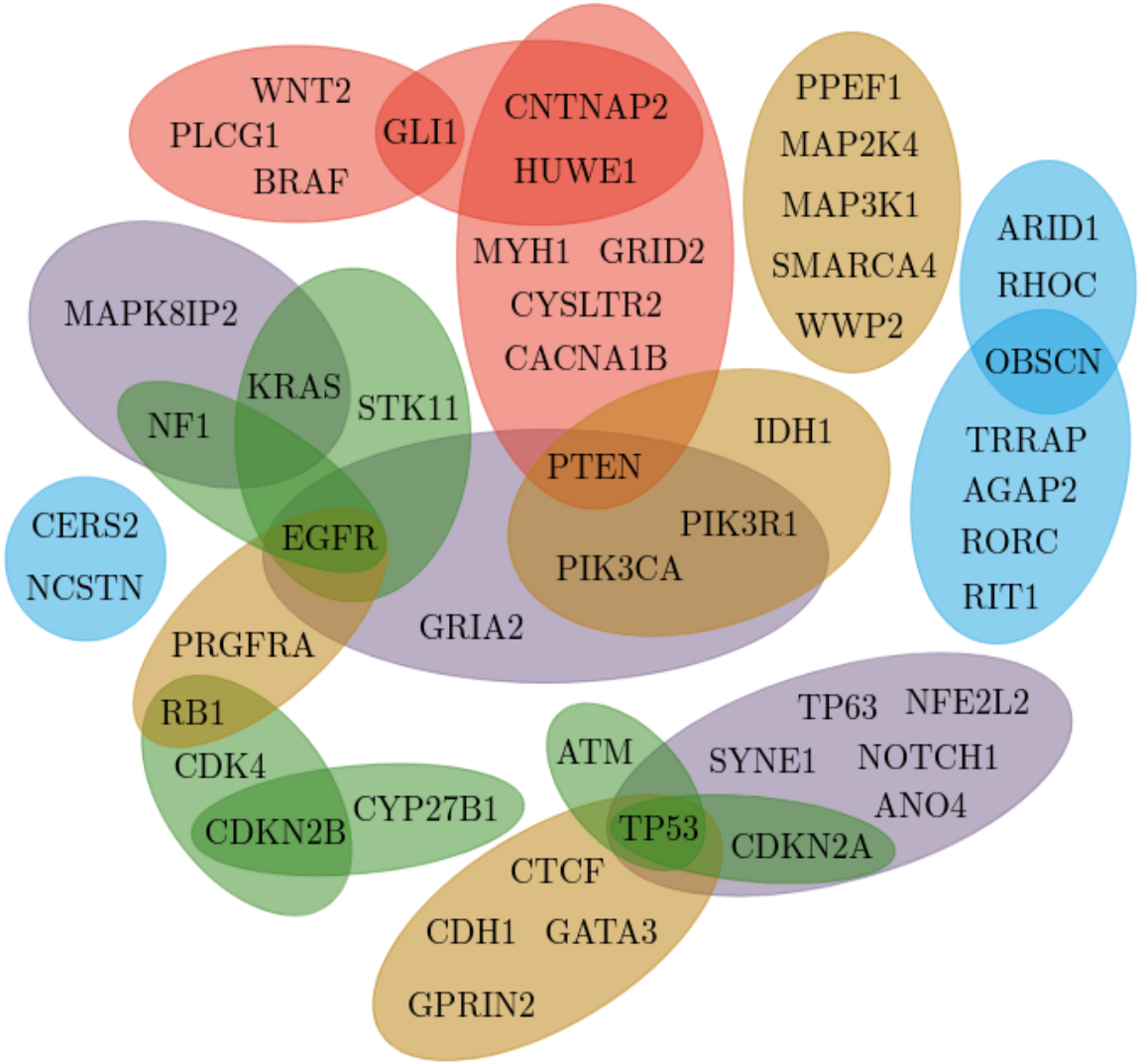
Un **grafo di geni** è un grafo completamente connesso in cui ad ogni arco sono assegnati due pesi.

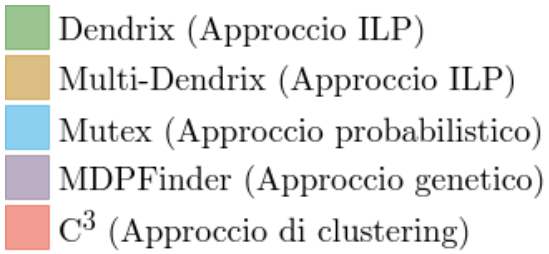


Il **peso negativo** di un arco (u,v) è il costo di posizionare u e v nello stesso cluster.

Il **peso positivo** di un arco (u,v) è il costo di posizionare u e v in cluster diversi.

Risultati


WNT2
PLCG1
BRAF
GLI1
CNTNAP2
HUWE1
MYH1
GRID2
CYSLTR2
CACNA1B
PPEF1
MAP2K4
MAP3K1
SMARCA4
WWP2
ARID1
RHOC
OBSCN
MAPK8IP2
KRAS
STK11
NF1
PTEN
IDH1
TRRAP
AGAP2
RORC
RIT1
CERS2
NCSTN
EGFR
PIK3R1
PIK3CA
GRIA2
PRGFRA
RB1
CDK4
CDKN2B
CYP27B1
ATM
TP63
NFE2L2
SYNE1
NOTCH1
ANO4
TP53
CDKN2A
CTCF
CDH1
GATA3
GPRIN2


Dendrix (Approccio ILP)
Multi-Dendrix (Approccio ILP)
Mutex (Approccio probabilistico)
MDPFinder (Approccio genetico)
C3 (Approccio di clustering)

Lavori futuri

L'identificazione dei pathway *driver* offre prospettive promettenti per migliorare l'efficacia delle terapie a bersaglio.

Lavori futuri

L'identificazione dei pathway *driver* offre prospettive promettenti per migliorare l'efficacia delle terapie a bersaglio.

Lavori futuri potrebbero integrare tecnologie emergenti, come il *single-cell sequencing*.

Servono algoritmi che considerino l'*eterogeneità tumorale* ed i meccanismi di resistenza adattativa alle terapie a bersaglio.

Grazie per l’attenzione

Problemi nel classificare le mutazioni

Per classificare le mutazioni è necessario valutarne la **funzione biologica**, ma questo rimane tutt'ora difficile da eseguire.

Attualmente è possibile categorizzare le mutazioni esaminandone la **frequenza**, poiché le mutazioni *driver* sono le più ricorrenti nei genomi dei pazienti.

Questo approccio funziona per geni frequentemente mutati, ma geni con relativamente poche mutazioni sono molto più comuni, e per questi analisi di frequenza **falliscono** nel classificarli.

Tipologia di approcci

Definizione. (Algoritmo *de novo*) Un algoritmo *de novo* identifica pattern utilizzando i soli dati genetici dei pazienti.

Osservazione. Gli algoritmi *de novo* potrebbero trovare risultati meno precisi.

Definizione. (Algoritmo *knowledge-based*) Un algoritmo *knowledge-based* integra la ricerca con dati esterni noti *a priori*.

Osservazione. Gli algoritmi *knowledge-based* sono limitati dall'attuale conoscenza incompleta dei pathway.

Approccio con ILP

Definizione. (Indicatrice di M) $I_M(j)$ è la variabile indicatrice che descrive l'insieme di geni M.

$$I_M(j) = 1 \iff j \in M$$

Definizione. (Indicatrice di $\Gamma(M)$) $C_i(M)$ è la variabile indicatrice che descrive quali pazienti copre M.

$$C_i(M) = 1 \iff \exists g \in M \mid i \in \Gamma(g)$$

Multiple Maximum Weight Submatrix Problem (MMWSP)

Definizione. (MMWSP) Data una matrice di mutazione A di dimensioni $m \times n$, ed un intero $t > 0$, si trovi la collezione $M = \{M_1, ..., M_t\}$ di sottomatrici colonna di A che massimizzi

$$W'(M) := \sum_{\rho=1}^{t} W(M_\rho)$$

	g_1	g_2	g_3	g_4
p_1	0	1	0	1
p_2	1	0	0	0
p_3	0	1	0	0
p_4	0	0	1	1